15 October 2004

Science

Vol. 306 No. 5695
Pages 357–556 $10

Cognition & Behavior

AAAS

Science

Special Issue

Cognition and Behavior

Volume 306
15 October 2004
Number 5695

A large number of brain areas are activated during higher cognitive functions. A special section in this issue of *Science* explores recent advances in research on cognition and behavior. [Image: Tim Rue/True Photo, modified from a suggestion by K. R. Ridderinkhof *et al.*]

For related online content, see page 367, or go to www.sciencemag.org/sciext/cognition

Departments

News of the Week

News Focus

398

419

Letters

Books *et al.*

Policy Forum

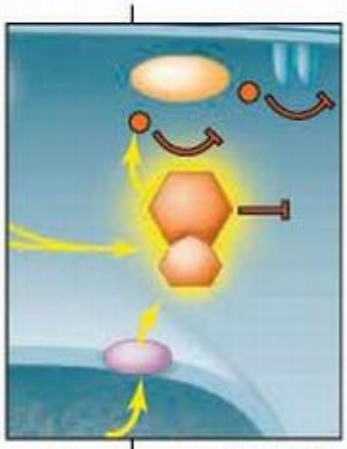
Perspectives

Contents continued

Science

Essay

Science Express www.sciencexpress.org



425 & 457

Technical Comment Abstracts

Brevia

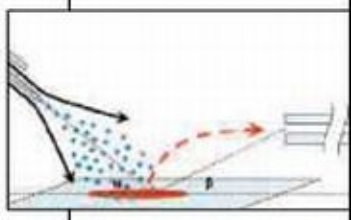
Research Article

Reports

423 & 466

Contents continued

Reports continued



388 & 471

479

AAAS
ADVANCING SCIENCE. SERVING SOCIETY

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals Mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2004 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership and subscription (51 issues): $130 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $500. Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery) $85. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #1254 88122. Publications Mail Agreement Number 1069624. Printed in the U.S.A.

Change of address: allow 4 weeks, giving old and new addresses and 8-digit account number. Postmaster: Send change of address to Science, P.O. Box 1811, Danbury, CT 06813–1811. Single copy sales: $10.00 per issue prepaid includes surface postage; bulk rates on request. Authorization to photocopy material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act is granted by AAAS to libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that $15.00 per article is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923. The identification code for Science is 0036-8075/03 $15.00. Science is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

Contents continued

Science
www.scienceonline.org

sciencenow www.sciencenow.org Daily News Coverage

Plankton Share the Spectrum
Diversity thrives when species use different wavelengths for photosynthesis.

How to Make a Uterus
Two genes implicated in the switch from egg laying to live birth.

What's in a Chimp's Toolbox?
Hidden cameras catch wild chimps using different tools for different tasks.

Trouble at the U.S.-Canada border.

science's next wave www.nextwave.org Career Resources for Young Scientists

Global/US: An Unexpected Career Move *J. Austin*
Border rules kept a UK citizen from returning to his U.S. lab after a weekend trip to Montreal.

Global/Europe: European Mobility Centres—Put Your 'Mobstacles' Aside *A. Kwint*
Most scientists love the international ambiance of the science world but face obstacles to mobility.

US: Tooling Up—The Paths of the Contrarian Job Seeker *D. Jensen*
Traditional scientists have taken different paths to finding nontraditional jobs.

UK: Science and the City *K. Arney*
Kat Arney hung up her postdoc boots to be a science information officer for a medical research charity.

MiSciNet: Motivating Young Children *E. Francisco*
A Native American of Acoma descent found she could inspire generations of young Indian scientists through teaching.

MiSciNet: Administrative Suggestions *M. P. DeWhyse*
College administrators can create a nurturing environment for underrepresented minority graduate students.

science's sage ke www.sageke.org Science of Aging Knowledge Environment

Perspective: Nuns and Monkeys—Investigating the Behavior of Our Oldest Old *J. A. Corr*
Studies of elderly people and nonhuman primates enhance each other. *related Cognition and Behavior section page 431*

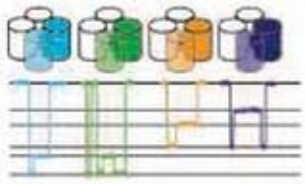
News Focus: Homing in on a Hormone *R. J. Davenport*
Fat-based molecules divert nematodes from survival mode.

News Focus: Wayward Sisters *M. Beckman*
Progestin siblings exert opposite effects on monkey hearts.

Monkeying around with aging.

science's stke www.stke.org Signal Transduction Knowledge Environment

Related Cognition and Behavior section page 431



NMDA receptor signatures.

Editorial Guide: Focus Issue—Signals Along a Road to the Mind? *E. M. Adler and N. R. Gough*
At the most basic level, signaling pathways govern learning, memory, and behavior.

Review: Role of Distinct NMDA Receptor Subtypes at Central Synapses *S. G. Cull-Candy and D. N. Leszkiewicz*
Distinct roles are emerging for NMDA receptor subtypes in synaptic plasticity, neuronal development, and pain.

Perspective: Coactivation of D1 and D2 Dopamine Receptors—In Marriage, a Case of His, Hers, and Theirs *A. Pollack*
Activation of a D1-D2 receptor complex stimulates a novel signaling pathway.

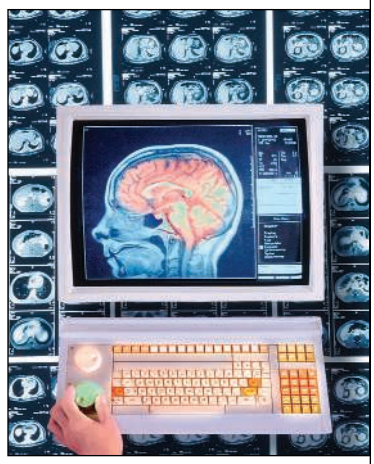
Teaching Resource: AMPA Receptor Cycling in the Synapse *A. Contractor and S. F. Heinemann*
This animation shows activity-dependent insertion of glutamate receptors in the postsynaptic membrane.

Separate individual or institutional subscriptions to these products may be required for full-text access.

Neuroscience and Neuroethics

Neuroethics, it appears, is a subject that has "arrived." The Dana Foundation is, for the second time since 2002, sponsoring a special lecture on this topic at this year's annual meeting of the Society for Neuroscience. AAAS, publisher of *Science*, also joined with Dana to produce a conference on "Neuroscience and the Law" earlier this year. The U.S. President's Council on Bioethics is now devoting serious attention to the topic. Companies are deploying functional magnetic resonance imaging (fMRI) to map brain activity as they assess the product preferences of prospective consumers (Coke or Pepsi?). There's even a new discipline called neuroeconomics. So something is going on here.

What got it started, and where is it headed? I think it emerged as new techniques and insights into human brain function gave us a dramatically revised notion of what might be possible. The first microelectrode recordings in active, behaving, nonhuman primates made it possible to look seriously at how valuation, choice, and expectation are encoded by single cells in particular parts of the brain. It further evolved with the development of fMRI and other noninvasive techniques for tracing neural activity in people. These studies are beginning to explain how particular brain structures are involved in higher functions (making difficult moral choices, for example) or in predisposing the individual to a particular kind of behavior.



In a different area, the successes of psychopharmacology in altering brain states and behavior have raised new problems of their own, not least in terms of how we may feel about the chemical manipulation of innate capacities. The list is long and ever growing: antidepressants, methylphenidate (Ritalin) for attention deficit hyperactivity disorder (ADHD), compounds that enhance alertness, and a new wave of drugs that may enhance memory formation and heighten cognitive ability.

Some of the questions now being raised by our expanded neuroscientific capacity are not exactly new. Consider, for example, the old issue of treatment versus enhancement. A child deficient in growth hormone could benefit from replacement therapy, and few would object to that, but its use by an aspiring teenage basketball player of normal height would raise questions. Now to the nervous system: Children with ADHD are often given methylphenidate after a physician considers their need. High school and college students without benefit of evaluation are using the same drug in the hope of improving their exam performance. Aside from the health risks associated with such drugs, what is it that bothers us here?

Perhaps it is our belief that the playing field should be level—we worry about the students who can't access the drug. Well, what about the kids who can't afford a preparatory course for taking a standardized test? Don't they raise the same questions about distributive justice? And suppose that we make the playing field level: All kids get the drugs, and all the sprinters get the steroids. Risks aside, are we comfortable with competition run in this way? Will the winners examine their enhanced selves and wonder "Was that really me?"

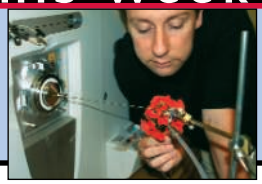
The ability to peer into brain processes also intensifies old privacy questions. Suppose that fMRI records become individually diagnostic with respect to some behavioral anomaly or predictive of some future tendency. Surely we would worry if they were used in insurance or employment contexts or in criminal litigation. Privacy protection would be guaranteed if the record were obtained as part of a medical procedure, but of course there are other possible sources. In the future, brain imaging techniques could conceivably be employed in the context of a court procedure as a test of truth-telling or subpoenaed in a case involving violence.

Finally, special issues arise when we penetrate into the philosophical territory where dualists and determinists debate over free will. As we learn more about the neurobiology of choice and decision, will we reach a point at which we feel less free? Perhaps more important for society, will we eventually know enough to change our view about individual responsibility for antisocial acts? There are those who worry about this. I am not among them, only because it seems so unlikely to me that our knowledge of the brain will deepen enough to fuse it with the mind. So, remaining convinced that my will is free, I am left to worry about the privacy of my inclinations and my thoughts.

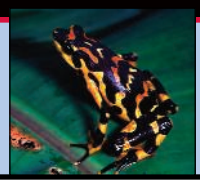
Donald Kennedy
Editor-in-Chief

Published by AAAS

This Week



PAGE 388
Mass spec on the move

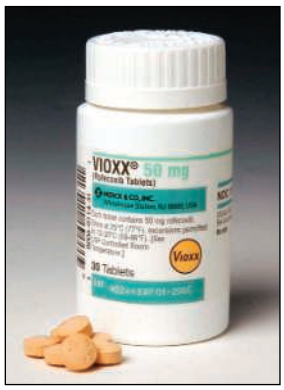


391
The global amphibian decline

DRUG SAFETY

Withdrawal of Vioxx Casts a Shadow Over COX-2 Inhibitors

On 30 September, the drug giant Merck announced that it was yanking its blockbuster anti-inflammation medicine, the COX-2 inhibitor Vioxx, off the market after an alarming pattern surfaced halfway through a 3-year colon polyp prevention study. Heart attacks and strokes had occurred at a much higher rate among the roughly 1300 volunteers on Vioxx (3.5%) than among the 1300 taking a placebo (1.9%). Within days, pharmacies were packing up their supplies of Vioxx and shipping them back to the company.



Reversal. From blockbuster to bust in 5 years.

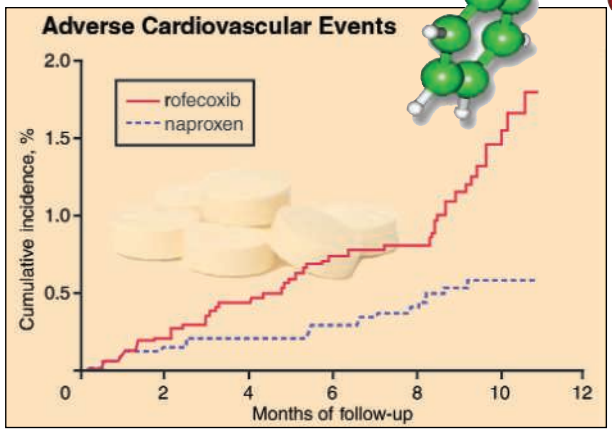
The scale of the withdrawal was unprecedented, casting a shadow over Merck, based in Whitehouse Station, New Jersey, and raising questions about the entire class of COX-2 inhibitors. Used primarily to treat arthritis and inflammatory pain, the drugs have earned billions of dollars since coming on the market more than 5 years ago. But the question could hardly be avoided: Did Vioxx collapse because of flaws unique to its chemistry, or would other COX-2 inhibitors suffer a similar fate?

"There are a lot of things we need to know now," says Garret FitzGerald, a pharmacologist and cardiologist at the University of Pennsylvania in Philadelphia. "The game has shifted."

Vioxx's propensity to trigger heart attacks and strokes isn't fully understood. But some experts believe that its valued mechanism—specifically, its ability to suppress a narrow set of molecules that mediate inflammation—may have been its downfall. Targeted drugs are all the rage, but many scientists worry that this particular targeting can upset a delicate balance that keeps blood-clotting at bay.

Drug regulators, among others, appear to be thinking along these lines. Last week, the European Medicines Agency in London said it would begin reviewing the safety of other COX-2 inhibitors, including Celebrex, made by Pfizer, based in New York City. U.S. experts at the Food and Drug Administration (FDA) and elsewhere cautioned against lumping other COX-2 inhibitors with Vioxx, but at the same time they have begun to review some studies of these drugs, including for pain, cancer inhibition, and Alzheimer's prevention. Richard Goldberg, chief of hematology and oncology at the University of North Carolina, Chapel Hill, learned for example that the National Cancer Institute and others overseeing his trial of Celebrex for preventing colon polyps, slated to enroll 1200 volunteers, were considering whether it might harm participants.

Manufacturers sought to reassure the public last week about their COX-2 products, a class that includes two Pfizer drugs on the market, Celebrex and Bextra, along with Prexige, made by the Swiss company Novartis, and Arcoxia, a Merck drug. The last two are approved in parts of Europe and are in late-stage development in the United States. Pfizer took a bold step, promoting claims of Celebrex's safety in full-page newspaper ads. But as some experts noted, studies of these drugs submitted to FDA did not last as long as the Vioxx colon polyp study. In that case, Merck didn't see serious problems until 18 months into the trial.



Molecule in trouble. Some experts say that even before the new data, a 2000 study showed that rofecoxib (Vioxx), compared here to naproxen, could cause cardiovascular problems.

Celebrex was the first COX-2 drug, introduced in early 1999. Before that, arthritis patients relied mainly on nonsteroidal anti-inflammatory drugs such as aspirin and naproxen (marketed as Aleve) to blunt symptoms. In some patients, though, these drugs can cause stomach problems.

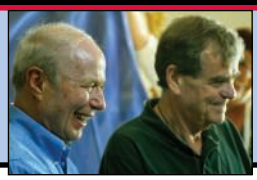
COX-2 inhibitors were hailed and heavily promoted as a major breakthrough because they home in on COX-2, an enzyme implicated in inflammation, while largely avoiding COX-1, which protects the stomach from gastric acids. Earlier anti-inflammatories had targeted both.

But preventing gastric upset may come at a cost. "Anyone who sits down with a pencil and paper and maps out the sequence of events" triggered by Vioxx "would have to say, 'Could this enhance thrombosis?' " says Benedict Lucchesi, a cardiovascular pharmacologist at the University of Michigan, Ann Arbor.

The theory Lucchesi favors, which FitzGerald also endorses, is based on how two fatty acids work. One, prostacyclin, stops platelet formation and prevents the cells from clumping; it also dilates blood vessels. The other, thromboxane, has the opposite effect, encouraging platelet clumping and constricting vessels. Anti-inflammatory drugs like naproxen suppress both prostacyclin, which plays a role in inflammation, and thromboxane. But COX-2 inhibitors block only prostacyclin; this may tilt the balance in favor of thromboxane and, potentially, blood clotting. So far the thrombosis theory has been supported only by animal studies.

Still, the COX-2 drugs on the market are unique molecules and differ in critical ways. For example, they vary in how tightly they target COX-2 and avoid COX-1. They also vary in how long they linger in the body. Even if the thrombosis theory holds, the risk of blood clots from COX-2 inhibitors almost certainly differs from drug to drug. Vioxx is both highly ▶

CREDITS (TOP TO BOTTOM): PHOTO AND STRUCTURE: COPYRIGHT MERCK & CO. INC., ALL RIGHTS RESERVED; GRAPH ADAPTED FROM D. MUKHERJEE *ET AL.*, *JAMA* **286**, 8 (2001)



targeted to COX-2 and has one of the longest half-lives, upward of 14 hours, a combination that some speculate may have triggered its problems. "We don't have a good explanation about why Vioxx is an outlier," says Eric Topol, who heads cardiovascular medicine at the Cleveland Clinic in Ohio and, along with some other physicians, has long harbored concerns about the drug. "It's always carried the worst risk of heart attack and stroke, of blood pressure elevation, of heart failure."

But some experts are not completely satisfied with the thrombosis theory. "I doubt that's the entire explanation" for Vioxx's dangerous effects, says Thomas Schnitzer, a rheumatologist and assistant dean of clinical research at Northwestern University in Chicago. Like all nonsteroidal anti-inflammatory drugs, Vioxx tends to boost blood pressure. Schnitzer wonders if this might be its Achilles' heel. Merck officials, however, told FDA that when they looked for a link between increased blood pressure and the heart attacks and strokes in the colon polyp trial, they didn't find one.

Topol, for one, believes more needs to be done: In an editorial released last week by *The New England Journal of Medicine*, he suggests that there could be "thousands of affected people" and calls for a congressional inquiry into Merck and FDA's handling of Vioxx in the years since it was approved.

–JENNIFER COUZIN

INFLUENZA

Crisis Underscores Fragility of Vaccine Production System

A snafu at a vaccine factory in Liverpool, U.K., has derailed U.S. plans to prepare for this year's flu season—and focused fresh attention on the fragile supply of essential vaccines.

Last week, Chiron, a pharmaceutical company based in Emeryville, California, announced that its Liverpool factory, which sells 90% of its vaccine to the United States, is unable to deliver any flu vaccine this year after British regulatory authorities effectively shut down the plant. The news sent U.S. authorities scrambling to ensure that the remaining vaccine supply—some 55 million doses, instead of the 100 million or more they had counted on—goes to those most at risk of complications and death, such as people over 65 years of age.

Chiron first reported on 26 August that its vaccines would be delayed because a small part of this year's batch of 50 million doses was contaminated with *Serratia marcescens*, a microbe that can cause opportunistic infections. Still, Chiron CEO Howard Pien assured a U.S. Senate Special Committee on 28 September that the company would eventually deliver 46 million to 48 million doses.

But on 5 October, the U.K.'s Medicines and Healthcare Products Regulatory Agency abruptly suspended Chiron's license to produce vaccines for 3 months, saying the company did not comply with so-called Good Manufacturing Practice regulations. Chiron, which acquired the plant last year when it bought the British company PowderJect, called the setback a "public health tragedy" but has declined to say how much of the vaccine is contaminated or what caused the problem. U.S. Food and Drug Administration (FDA) officials visited the plant in Liverpool last weekend to investigate. At a House hearing last week, acting director Lester Crawford appeared pessimistic that part of the batch might be salvaged.

The shortage comes at a time when a record 185 million Americans were advised to get flu shots. In guidelines issued this spring, the Advisory Committee on Immunization Practices (ACIP) had added children between 6 and 23 months and their close contacts to the list of groups that should get the vaccine. (It already included people over 50, patients with chronic illnesses, pregnant women, and nursing-home residents, as well as anyone who might transmit the virus to people in these groups.) After the Chiron announcement, ACIP pared down the list during a hastily convened meeting that same day, striking, for instance, parents of young children and healthy people between 50 and 65 and urging anyone not in a risk group to forgo the shots this year.

The number of people actually vaccinated is always much smaller than the recommended numbers, says immunologist Paul Offit of the Children's Hospital of Philadelphia, a former member of ACIP. Even so, the remaining lots—about 54 million doses of injected, killed vaccine from Aventis Pasteur, and 1 million to 2 million of FluMist, a live intranasal vaccine produced by MedImmune—will not be enough, Offit predicts: "There will be people who want the vaccine, who can't get it, and who will die because of that."

Underlying the problem is an exodus of pharmaceutical companies from the vaccine business, which is widely seen as risky and not lucrative. The dwindling manufacturing base has led to previous severe shortages of some vaccines in the United States (*Science*, 15 March 2002, p. 1998). Production of the flu vaccine is especially vulnerable because its exact composition changes annually. Companies produce the vaccine between March and September every year in a tightly choreographed process. That's why no company can easily fill the gap left by Chiron, says David Fedson, a former medical director of Aventis Pasteur MSD who lives in Sergy Haut, France. The fragile supply could prove catastrophic should a new pandemic flu emerge, Fedson cautions (see p. 394).

First in line. New interim recommendations give priority to members of high-risk groups like those over 65.

Many solutions have been floated—from subsidizing companies to building a new government-operated vaccine plant—but little has been done. The current crisis should put the issue back on the agenda, says epidemiologist Arnold Monto of the University of Michigan, Ann Arbor: "We really need a sea change."

–MARTIN ENSERINK

CREDIT: NICK UT/AP PHOTO

Published by AAAS

MEDICINE

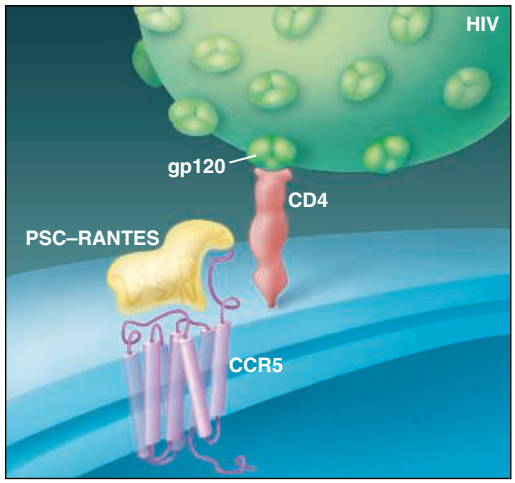
Microbicide Shuts the Door on HIV

Microbicides have long had a stepchild status in the AIDS research community. Industry has had little interest in developing a topical gel or cream that can stop HIV at the vagina or rectum, and the products that have moved furthest in human studies are soaps and other substances that do not specifically target the virus. But over the past few years, nonprofits and governments have poured substantial money into microbicide research and development, bringing forward several cutting-edge concepts. On page 485 of this issue, an international team of researchers describes a monkey study that features one such strategy: a microbicide specifically designed to block HIV's ability to infect its favorite target cell. "They are applying true antiretroviral science to microbicides," says Mark Mitchnick, who heads R&D for the nonprofit International Partnership for Microbicides in Silver Spring, Maryland.

HIV typically establishes an infection by first attaching to CD4 receptors on white blood cells and then grabbing a second receptor known as CCR5, which normally responds to immune chemicals called chemokines. In the study, clinical immunologist Michael Lederman of Case Western University in Cleveland, Ohio, teamed up with Oliver Hartley of the University of Geneva in Switzerland, whose lab had created a CCR5 inhibitor, PSC-RANTES, by modifying one of the chemokines that uses the receptor. Working with a group led by Ronald Veazey of the Tulane National Primate Research Center in Covington, Louisiana, they applied different doses of the compound to the vaginas of 30 monkeys. Fifteen minutes later, they challenged the animals with an intravaginal dose of a chimeric monkey/human AIDS virus. In animals given relatively high doses of PSC-RANTES, 12 of 15 completely resisted infection. "This is the first paper that says if you target the susceptible cells, you can block infection by mucosal cells," says Robin Shattuck of St. George's Hospital Medical School in London.

Many mysteries remain about the mechanism of sexual transmission of HIV, and Lederman suggests that this study may help clear up a critical one. Although other studies have shown sexual transmission of the virus through routes that don't involve the CD4/CCR5 nexus, "this experiment suggests that blocking CCR5 is enough to prevent infection," says Lederman.

Yet he is quick to point out that the dose of PSC-RANTES required for protection in this study is "too high to be practical." Manufacturing the amount of PSC-RANTES needed to protect each monkey proved extremely expensive, so the Geneva team is now attempting to develop a cheaper version of the molecule. Lederman and others also note that several companies have developed potentially cheaper, small-molecule CCR5 inhibitors. Veazey, working with AIDS immunologist John Moore of Cornell University's Weill Medical College in New York City, last year found that one of these protected two of 11 monkeys in a viral challenge experiment. "We've done better since," says Moore.



Blocked dock. PSC-RANTES prevents infection of CD4 cells by blocking HIV's gp120 from binding to CCR5.

Lederman and colleagues also raise the possibility that their study may have set the bar too high; the monkeys were given hormones to make them more susceptible to the virus. Smaller amounts of PSC-RANTES might therefore work in the real world. Some human studies have shown that the transmission of HIV from male to female may occur as infrequently as one out of every 2000 sexual encounters. But a group led by Christopher Pilcher of the University of North Carolina, Chapel Hill, published a study in the May issue of the *Journal of Infectious Diseases* reporting that males in the initial stage of an HIV infection can transmit as frequently as once out of every four encounters.

Shattuck says it should be assumed that a microbicide will have to protect against high-dose challenges. Still, he is heartened by the new study. "We've moved from an era of trying unsophisticated approaches to rational drugs that we understand," Shattuck says. "It's a new phase in microbicide approaches."

–JON COHEN

ILLUSTRATION: K. SUTLIFF/*SCIENCE*

ScienceScope

$60 Million Imaging Initiative to Track Alzheimer's

A 5-year, $60 million public-private research project launched this week will explore whether brain imaging can be used to track the development of early Alzheimer's disease.

The Alzheimer's Disease Neuroimaging Initiative will follow up on small studies suggesting that magnetic resonance imaging and positron emission tomography can be used to forecast when individuals with early signs of memory loss will develop Alzheimer's. The National Institute on Aging (NIA) and other federal sponsors are putting up about two-thirds of the money; the rest will come from drug companies and nonprofit groups. Fifty sites will enroll 800 adults, some with no signs of disease, some with mild cognitive impairment, and some with early Alzheimer's, and track them for up to 3 years. The lead investigator is Michael W. Weiner of the Department of Veterans Affairs and the University of California, San Francisco.

The study is meant to collect baseline data—not test treatments—although some patients will likely be taking Alzheimer's medications, says NIA neuroscientist Neil Buckholtz. NIA director Richard Hodes hopes the initiative will be a "landmark study." – JOCELYN KAISER

CITES Cuts Caviar Exports

A United Nations conservation agency has cut exports of caviar (sturgeon eggs) from the Caspian Sea region. But last week's move by the 166-member Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) disappointed environmentalists, who say the agency backed away from doing more to protect sturgeon stocks, which have dwindled by as much as 90% in some areas.

CITES says five Caspian nations—Russia, Iran, Kazakhstan, Azerbaijan, and Turkmenistan—can export 113 metric tons of caviar, down 20% from last year. But CITES said next year's quota could be bigger if the nations make greater efforts to control poachers, who experts say produce up to five times more caviar than legal fishers. Earlier, CITES officials had threatened to bar exports unless nations did more to document and control poaching (*Science*, 10 September, p. 1547).

"CITES has flip-flopped under pressure by Caspian states and the caviar industry," says Vikki Spruill of SeaWeb, a conservation group based in Washington, D.C. But CITES deputy secretary Jim Armstrong says, "the new approach ... gives the governments a strong economic stake in tackling illegal fishing." –CHRISTOPHER PALA

Published by AAAS

CHEMISTRY

Mass Spectrometrists Salivate Over Recipe for Ions Alfresco

A new way of making ions could revolutionize the venerable practice of mass spectrometry, in which ionized molecules are identified by their weight. Standard ionization techniques work only within cumbersome vacuum chambers or require specially prepared samples. But a simple spritz from a gas jet can liberate ions from almost any surface, even in the presence of air, a team of analytic chemists reports on page 471 of this issue of *Science*. That means researchers can analyze a vast variety of samples simply by holding them under the jet. The technique could be used in airports to "sniff" luggage for traces of explosives, in orchards to test fruit for pesticide residues, and in many other venues outside the laboratory.

Blooming simple. Lead author Zoltán Takáts demonstrates new technique for wafting ions into a spectrometer (left).

"It's the greatest thing since night baseball," says John Fenn, a chemist at Virginia Commonwealth University in Richmond. Fenn won a share of the 2002 Nobel Prize in chemistry for developing a technique on which the new method is based. Gary Van Berkel, a mass spectrometrist at Oak Ridge National Laboratory in Tennessee, says the technique has a wealth of potential applications. "My mind's been racing since I read the abstract," he says. "I came in this morning and set up an experiment, and in 5 minutes I had it working."

Dubbed desorption electrospray ionization (DESI), the new method combines elements of other well-established techniques, report Zoltán Takáts, R. Graham Cooks, and colleagues at Purdue University in West Lafayette, Indiana. Researchers can ionize large molecules by dissolving them in a solvent and using an intense electric field to pull tiny charged droplets of solution from the end of a needle—an approach known as electrospray ionization, for which Fenn won the Nobel Prize. The new technique uses an electrospray jet differently, to shoot ionized droplets of solvent at a sample.

In that regard, DESI resembles techniques in which beams of other ions or laser light blast ions out of a sample's surface. However, the ion beam technique works only in a vacuum chamber, and laser samples usually must be specially prepared and must fit into the laser rig. DESI works with everyday surfaces and sucks ions into the spectrometer through a sampling tube. Using the method, Takáts and Cooks have detected traces of the explosive RDX on a leather surface and residue of the chemical weapon DMMP on a nitrile glove; tracked organic compounds in seeds and stems of plants; and even sniffed out an antihistamine on the skin of a person who had taken the drug 40 minutes earlier. The team has patented the technique, and a small start-up company will try to commercialize it.

Van Berkel suspects that DESI will prove most useful for analyzing laboratory samples, such as the plates generated in gel electrophoresis measurements. But Albert Heck, a mass spectrometrist at Utrecht University in the Netherlands, says the technique opens the way for taking mass spectrometers out into the world and analyzing surfaces wherever they may be found. As they travel down life's road, mass spectrometrists can now stop and ionize the roses.

–ADRIAN CHO

CREDIT: Z. TAKÁTS *ET AL.*

GRADUATE EDUCATION

Hughes, NIH Team Up on Novel Training Program

The country's biggest private sponsor of biomedical research is joining hands with the National Institutes of Health (NIH) in an unusual arrangement to train interdisciplinary scientists.

Under the initiative, the Howard Hughes Medical Institute (HHMI) will provide up to $1 million over 3 years to each of 10 institutions to help them create Ph.D. programs that integrate biomedicine with the physical sciences and engineering. The money will go toward hiring staff and developing curricula. Once the programs are up and running, the National Institute of Biomedical Imaging and Bioengineering (NIBIB) will provide 5 years of funding to support the actual training of the graduate students. The total cost of the initiative is estimated at $35 million.

The 4-year-old NIBIB already funds training programs at 21 schools around the country. What is unusual about this effort, however, is that HHMI—not NIBIB—will choose the participating institutions. After 3 years Hughes will hand the program over to NIBIB, which will review an institution's progress before providing additional funding. "Although phase II funding is not guaranteed, we expect that all the programs will do well enough to qualify," says NIBIB's Henry Khachaturian. Each program is expected to train up to 10 students.

HHMI officials approached NIBIB with the idea to "ensure sustainability of the programs that we would be helping to create," says Peter Bruns, HHMI's vice president for grants and special programs. "It's unrealistic to start a training program without making sure that students will have continued funding." NIBIB welcomed the opportunity "to foster interdisciplinary training in a planned way," says institute director Roderic Pettigrew. "HHMI is better equipped than NIH to underwrite and develop the infrastructure for new programs. NIH, on the other hand, is well equipped to support programs that are fully established."

Although observers like the idea of pooling public and private resources for graduate training, some wonder about the wisdom of having a private foundation, in effect, select grantees for a federally funded program. "If the institutions chosen by HHMI are really the cream of the crop, why do they need a protected competition for funding from NIBIB?" asks one society official, who requested anonymity. A good approach might be "for HHMI and NIBIB to work together on all aspects of selection and administration from day one," says Peter Katona, director of the Whitaker Foundation, a major supporter of research training in biomedical engineering.

NIBIB officials say the agency will help HHMI select appropriate reviewers and ensure that a majority of them will be available for reviewing phase II applications. Guidelines for the competition, open to any U.S. institution granting Ph.D.s in biology, are online at www.hhmi.org/grants/pdf/comp_annc/2005_nibib_program.pdf.

–YUDHIJIT BHATTACHARJEE

Published by AAAS

GENETICS

Disease Backs Cancer Origin Theory

Almost all cancer cells have gained or lost entire chromosomes. Despite the genetic turmoil this causes, scientists have disagreed for nearly a century about whether this abnormality and other types of genomic instability, such as that caused by DNA repair defects, are the starting gun for cancer or merely a result of it. A study published online in *Nature Genetics* this week provides the strongest evidence yet for the starting gun theory by showing that mutations in a gene involved in ensuring proper chromosome number result in childhood cancer.

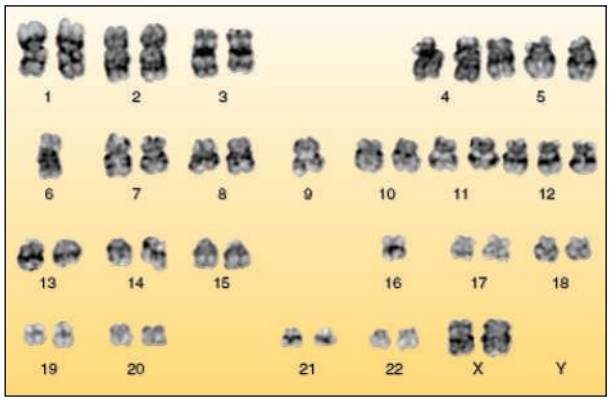
"The connection between chromosomal instability and cancer is now unassailable," says Bert Vogelstein, an oncologist at Johns Hopkins University School of Medicine in Baltimore, Maryland. "This study will stimulate a lot of research into whether mutations in genes [involved in chromosome maintenance] contribute to other types of cancer."

In 1914, German biologist Theodor Boveri noticed that the cancer cells he was studying contained an abnormal number of chromosomes, a state called aneuploidy. The observation led him to postulate that the condition was a root cause of cancer. But as researchers began to discover that mutations in specific oncogenes and tumor-suppressor genes were enough to set cancer in motion, the aneuploidy theory fell out of fashion. Now it's back, thanks to a series of studies in the mid-1990s on the larger issue of genomic instability. For example, Vogelstein and others showed that mutations in genes required for DNA repair led to a hereditary form of colon cancer, indicating that the destabilization of a cell's genome could instigate cancer. But the field is still deeply divided between scientists who believe genomic instability must happen early for cancer and those who say it happens later and may not even be required.

In the new study, a team led by cancer geneticist Nazneen Rahman of the Institute of Cancer Research in Sutton, U.K., screened the DNA of eight families with mosaic variegated aneuploidy (MVA)—a genetic disorder in which more than 25% of a patient's cells are aneuploid and childhood cancers such as rhabdomyosarcoma and leukemia occur much more frequently than normal. In five of these families, the group identified a child with mutations in both copies of a gene called *BUB1B*. All five children had a high percentage of aneuploid cells, and two have already developed cancer. The gene found mutated in these children encodes a protein previously shown to help guarantee that the right number of chromosomes are passed from cell to cell. The new work is the first to show that defects in *BUB1B* or any other genes guiding a cell's chromosome partitioning system lead to a human disorder.

"This indicates that aneuploidy has a direct causal role in cancer," says Rahman. Moreover, she says, the fact that a genomic instability like aneuploidy arises early in the life of someone with MVA argues that it is an incipient event in the disorder's cancer development and not a side effect of other processes. "This study will be a major part of the armory for people who argue that aneuploidy is a cause, not an effect, of cancer," contends Rahman.

Just because early genomic instability leads to cancer in MVA doesn't mean it's the trigger in all cases, says William Dove, a geneticist at the University of Wisconsin, Madison. His group has been unable to detect this process in a mouse model of intestinal cancer.



Wrong number. The abnormal number of chromosomes seen in this child may give a clue to the origins of cancer.

"Rahman's study provides very important evidence that early aneuploidy *can* cause cancer," he says, "but it doesn't close the debate."

Vogelstein agrees that other cancers should be studied. Unlike the tumors arising from MVA, he says, most cancers are not hereditary. "So it still leaves the door open as to whether this applies to [spontaneous] cancers, … but this is a giant step forward for those who believe early instability predisposes to cancer."

Establishing an accurate timeline for cancer progression should help researchers develop therapies targeted at preventing and treating the disease. Says Dove, "If we know the nature of the enemy, we will have a better way of attacking it."

–DAVID GRIMM

CREDIT: SANDRA HANKS

ScienceScope

Voters Warm to California Stem Cell Measure

A new poll suggests that voter support for California's Proposition 71 is strengthening. But a few key organizations have decided not to back the measure, which would issue $3 billion in state bonds to fund human embryonic stem cell work in the state (*Science*, 10 September, p. 1544).

A Field Poll of 549 likely voters taken at the end of September showed 46% in favor of the measure, with 39% opposed. (The poll had a 3.5% margin of error.) That's up from a near tie in an earlier poll, and pollsters found that voters familiar with the measure supported it by a wider (58–34) margin, suggesting that a multimillion-dollar ad campaign by backers is paying off.

But two influential groups have declined to endorse the measure, citing cost concerns. The San Mateo County Medical Association late last month withdrew an earlier endorsement, officially becoming neutral. And the San Francisco Bay Area's largest biotechnology industry association, BayBio, also opted not to take a position. BayBio "supports the elimination of all federal restrictions currently limiting stem cell research," it said in a statement. But group president Matt Gardner told reporters that some members of BayBio's board worried that the bonds could saddle California with debt and prevent future tax breaks for companies.

–GRETCHEN VOGEL AND DAVID MALAKOFF

No Meeting of the Minds at Harvard on Women Faculty

They may have broken bread together last week, but Harvard faculty members didn't much enjoy their conversation with President Lawrence Summers and Dean William Kirby over declining numbers of women being offered tenure. "Their reaction was like that of an elephant that's been bitten by a mosquito," says a biologist, one of 50 women at the 6 October luncheon arranged after the group aired its complaints (*Science*, 17 September, p. 1692). Summers took a decidedly anti–affirmative action stance at the meeting, says the participant, who requested anonymity, telling the group that "Harvard could not make hires based on anything other than pure merit."

The women have formed a Senior Faculty Caucus for Gender Equality to press their case for more competitive salaries and the inclusion of at least one woman on departmental search committees. Kirby says he will soon be writing to the faculty on "how I believe we can best search for a talented and diverse faculty."

–YUDHIJIT BHATTACHARJEE

Published by AAAS

ECOLOGY

Global Survey Documents Puzzling Decline of Amphibians

Almost a third of the world's amphibians are threatened with extinction, according to the first global survey of the situation. And it's not clear what's killing many of them off. "It's very sobering," comments David Wake of the University of California, Berkeley, about the assessment, described in a paper published online by *Science* this week (www.sciencemag.org/cgi/content/abstract/1103538).

Scientists first noticed the perilous state of many amphibians in the late 1980s. Many common species were becoming hard to find, even in national parks and other protected areas. In addition to a loss of habitat, studies pointed to herbicides, stronger ultraviolet light, and a fungal disease called chytridiomycosis. There was also speculation about the role of climate change and invasive species. Despite an accumulating stack of evidence, there was no global picture of all 5743 known species.

The $1.5 million Global Amphibian Assessment project, funded by several federal and nongovernmental donors, was launched in 2001 to provide that global picture. Simon Stuart of the International Union for Conservation of Nature and Natural Resources (IUCN) and colleagues at Conservation International and NatureServe, a biodiversity clearinghouse, began by dividing the world into 34 regions. They assigned a herpetologist to assemble a species list for each region and seek out information such as trends in abundance, distribution, and threats. More than 500 herpetologists reviewed the data. "The effort is unprecedented," says Michael Lannoo of Ball State University in Muncie, Indiana.

The next step was to evaluate the chance that each species would go extinct, according to IUCN "Red List" criteria. Not only are a third threatened, they found, but 7.4% of all amphibians—427 species—qualify for the highest IUCN threat level, known as critically endangered. Moreover, both figures are certainly underestimates, Stuart says, because too little is known about 1294 rare species to gauge their status. Stuart is seeking funding that would allow his team to update the database frequently and review it completely every 3 years.

The survey attempted to chart trends in amphibian species as well. One approach was to ask the expert reviewers what was happening to populations. Some 43% of amphibian species are dwindling in numbers, they reported; 27% are stable, and fewer than 1% are increasing. The status of the rest is unknown.

Another method was to look at species for which data existed in 1980—when declines apparently began—and compare their Red List status, then and now. The situation has gotten worse over the past 2 decades for 435 species, the survey reveals. (Again, this is likely an underestimate, Stuart cautions, because the decline of many species could have gone undetected.) In North America and Europe, the reason is largely habitat loss, whereas in East Asia it is humans hunting for food. But there is no obvious cause for the declines in the Neotropics and Australia, which host the majority of rapidly declining species.

Disappearing. Like many amphibians, the harlequin toad (*Atelopus varius*) is in serious decline for unknown reasons.

"The bottom line is that there's almost no evidence of recovery and no known techniques for saving mysteriously declining species in the wild," Stuart says. "It leaves conservation biologists in a quandary."

—ERIK STOKSTAD

NOBEL PEACE PRIZE

Kenya's Maathai Wins for Reforestation Work

Arrested, beaten, and jailed for her efforts, environmentalist and political activist Wangari Maathai of Kenya has won the 2004 Nobel Peace Prize.

Maathai, 64, is the first African woman to win the prize, announced last week, and the first to be honored for environmental work. The founder of the Green Belt Movement, which since 1976 has organized local groups to plant an estimated 30 million trees across eastern and southern Africa, Maathai was a longtime opponent of Kenya's former strongman Daniel arap Moi. She was physically attacked by opponents on several occasions and was once released from jail only after Amnesty International helped fuel international protests.

Seeds of change. Maathai's tree-planting program has attracted global attention.

Since 2002 she has served as deputy environment minister under President Mwai Kibaki and also holds a seat in Kenya's parliament.

In awarding the prize, the Norwegian Nobel committee said Maathai "combines science, social commitment, and active politics. More than simply protecting the existing environment, her strategy is to secure and strengthen the very basis for ecologically sustainable development."

Maathai's accomplishment also breaks new ground by recognizing environmental activism as worthy of a prize normally awarded for peacemaking and human-rights advocacy. "Peace depends on our ability to secure our environment," said Ole Danbolt Mjoes, the Nobel Committee chair.

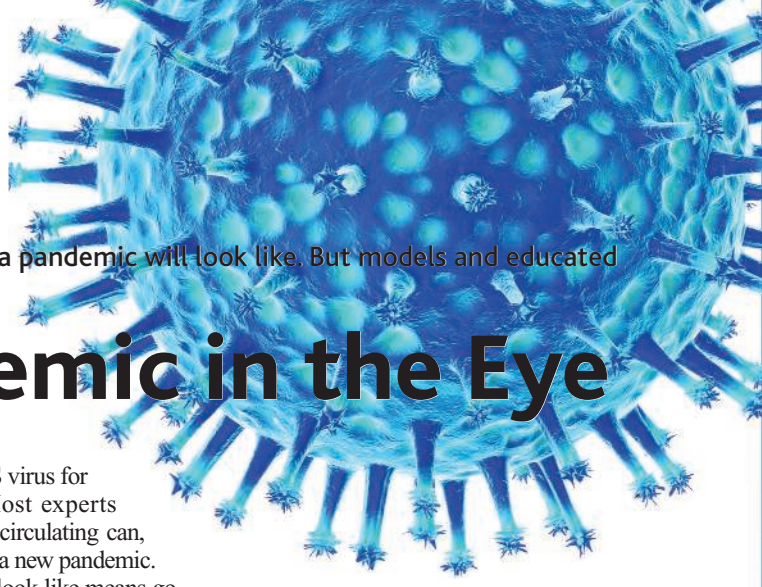
Maathai earned a Ph.D. from the University of Nairobi, one of the first women in the region to do so. She later chaired the school's department of veterinary anatomy, also a first for a woman. Maathai is "delightful, ebullient, and dynamic," as well as a keen thinker, says Chad Oliver of the Yale School of Forestry and Environmental Studies in New Haven, Connecticut, where Maathai was a visiting scholar in 2002. "She's able to look at a cloud of information and cut right through to the core."

Since winning the award, Maathai has provoked controversy by restating her belief that scientists may have created the HIV virus to harm Africans. Many prominent Africans have endorsed that fringe idea because the epidemic has hit the continent exceptionally hard, says Samuel Kalibala of the International AIDS Vaccine Initiative in Nairobi. But Maathai's remarks are unfortunate, he says: "We should not be diverted from fighting AIDS by trying to blame others."

—GRETCHEN VOGEL AND DAVID MALAKOFF

CREDITS (TOP TO BOTTOM): ROBERT PUSCHENDORF; KAREL PRINSLOO/AP PHOTO

Published by AAAS



Researchers have no way of knowing what the next influenza pandemic will look like. But models and educated guesses are disconcerting

Looking the Pandemic in the Eye

Ask flu experts about their worst nightmare and they may tell you something like this. Somewhere in Asia, a new flu virus is born that's able to jump from one human to the next, yet is cloaked in avian proteins that human immune systems have never seen before. Laying low at first, the virus sickens and kills a small number of people, while it's getting better at the human-to-human transmission game. When authorities finally notice the expanding cluster of flu cases, the virus has already moved on. It takes advantage of flights that connect Asia's major cities to the rest of the world, popping up simultaneously in Sydney, Los Angeles, and London.

Hundreds begin to die, literally drowning as fluid fills their lungs. A stunned public demands a vaccine, drugs—anything—but no vaccine will be available for months, and antivirals are in short supply; the question is, who gets them? Panic and riots erupt while schools, businesses, and transportation systems are shutting down. Overcrowded hospitals start turning away desperate patients. There aren't nearly enough doctors and nurses to take care of the sick and dying, nor enough coffins. When the outbreak finally peters out 18 months later, more than 2 billion people have become ill, and more than 40 million are dead—twice the number claimed by AIDS in 25 years.

True, that's a worst-case scenario—but few experts dismiss it out of hand. After years of neglect, the threat of a new pandemic is back on the world's radar screen, beeping noisily. Public health experts, virologists, and disease modelers are struggling to envisage how fast it would spread, how many it would kill, what it would cost, and most of all, how best to fight it.

The efforts were spurred in part by severe acute respiratory syndrome (SARS), the planet's close brush with pandemic disaster last year. The SARS virus wasn't all that contagious, striking fewer than 9000 people before it was brought under control. But the world may not be so lucky next time. Nor does it take a newcomer like the SARS virus for a pandemic to occur. Most experts agree that flu strains now circulating can, and eventually will, spawn a new pandemic.

Predicting what it will look like means going out on a limb, however, because everything depends on which flu strain is the culprit and how virulent it is—two questions no one can answer. Still, researchers can crunch the numbers for a range of assumptions. They end up with a series of scenarios—from something quite benign to an "overwhelming and potentially catastrophic event," says Martin Meltzer, an economist and disease modeler at the U.S. Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia.

Put to bed. A ballroom was turned into an emergency infirmary at the University of Massachusetts during the 1957 "Asian flu" pandemic.

Even trickier to predict are a pandemic's social, political, and economic fallout. "Go ask the fiction writers what could happen," Meltzer says. It seems certain, though, that a pandemic will raise agonizing dilemmas about who should be first to receive drugs, vaccines, and medical care—an issue that most countries haven't even begun to debate.

Virgin territory

Flu pandemics occur when a new virus emerges that's easily transmissible between people and also finds virgin territory in the human population because no one is immune. This happens when one or both of the virus's envelope proteins (hemagglutinin and neuraminidase, the H and N in names like H5N1) have never before circulated in humans.

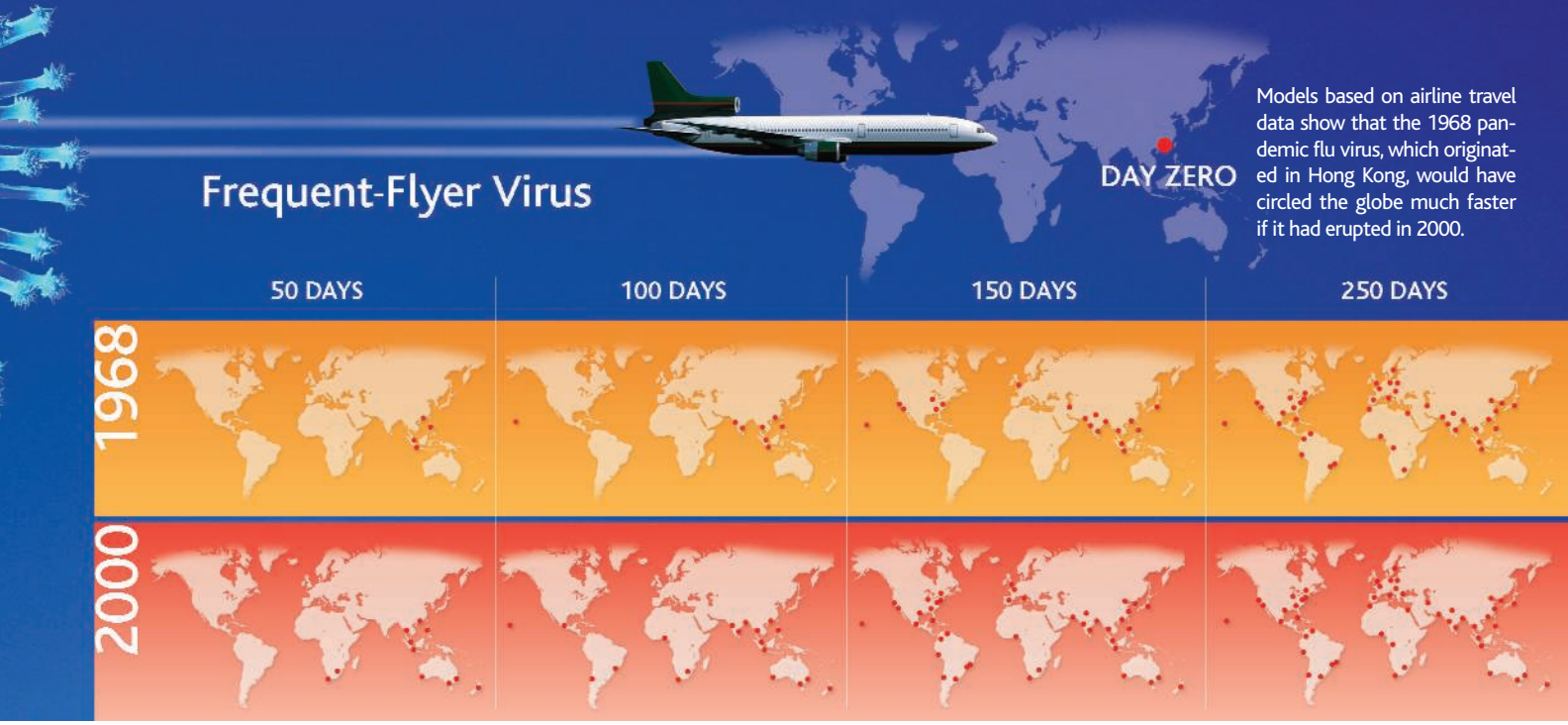
By far the most terrifying example is the 1918–19 "Spanish flu" pandemic, during which at least 20 million people, and perhaps as many as 100 million, are believed to have perished. Most of that virus's genetic baggage has been reassembled from preserved tissue scraps and an Alaska victim's frozen body. In a paper published in last week's issue of *Nature*, researchers reported that a modern flu strain equipped with the 1918 hemagglutinin is highly pathogenic to mice—a finding that may help clarify why the 1918 virus was so deadly. It's still unclear where the virus came from, however; nor are researchers sure about the origins of two subsequent, milder pandemics that struck in 1957 and 1968.

For decades, the dominant theory was that new pandemic viruses arise when avian and human flu viruses reassort, or hybridize, inside pigs, which can be infected with both. (Chinese farms, where ducks, humans, and pigs mingle, were seen as plausible locales.) But since 1997, three avian flu viruses—including H5N1, the virus that has infected poultry in 10 Asian countries—have been found to infect humans directly. Now, the predominant worry is that humans infected with both avian and human viruses may be mixing vessels.

Fortunately, chances of this happening still seem low, says Neil Ferguson, an epidemiologist at Imperial College in London. Even if you assume that reassortment occurs in each and every patient infected with the two viruses—which is unlikely—more than 600 people would have to be infected with H5N1 to create a 50% chance of reassortment, Ferguson and his colleagues wrote ear-

CREDITS (TOP TO BOTTOM): MATTHIAS KULKA/CORBIS; BETTMANN/CORBIS

Published by AAAS



lier this year in *Science* (14 May, p. 968). So far, fewer than 50 people in Vietnam and Thailand are confirmed to have been infected with H5N1. What's more, most reassortants are likely to pose no threat.

Assuming a new pandemic virus emerges, how might it behave? Epidemics can be modeled several ways, but mathematicians always need a number of key parameters, such as the basic reproductive number (R_0), which denotes the number of secondary infections resulting from one patient, the attack rate (the percentage of people who get sick after being exposed to the virus), the chance of becoming infected when in close contact with a patient, the incubation period, and the mortality rate.

For many diseases, those variables are reasonably well known and more or less constant. Not pandemic flu; even year-to-year changes in the influenza virus make for difficult modeling, says Ira Longini, an expert at Emory University in Atlanta—which is why modelers have tended to stay away from flu.

But faced with what many perceive as a gathering threat and using past pandemics as a rough guide, modelers are beginning to tackle the problem. The Models of Infectious Disease Agent Study (MIDAS), for instance, a network funded by the U.S. National Institutes of Health that includes Longini's group, this summer made work on flu pandemics its top priority. The U.S. government is keenly interested in the results, Longini says, because models can help decide how best to deploy drugs and vaccines.

The models all suggest that pandemic flu is unlikely to be contained using the old-fashioned public health measures that put the SARS genie back into the bottle, such as isolating patients and tracing and quarantining contacts. SARS has an incubation period of about 6 days during which infected people don't seem to infect others—precious time health authorities could use to trace those exposed but still healthy. With flu, they'd have only about 2 days on average. Moreover, SARS's severe symptoms helped identify patients, whereas flu can be as mild as the sniffles.

The only exception may be very early on, notes Ferguson. When the virus is still struggling to replicate among humans, surveillance and quarantine, perhaps helped by aggressive use of antiviral drugs, might nip a pandemic in the bud—which is why the World Health Organization is exploring a plan to ship antivirals to the cradle of a potential pandemic (see p. 394).

Once a virus was on the loose, jumbo jets would likely spread pandemic flu faster than ever in history. In a model published last year, Rebecca Grais and her colleagues at Johns Hopkins University in Baltimore, Maryland, collected data on the number of passengers traveling daily among 52 major cities around the globe and then calculated how fast the 1968 strain would have spread had it surfaced in 2000. Although the model has its limits, the trend is clear: The outbreak would peak in most of the 52 cities within 6 months (see graphic above). In the same model fed with travel data from 1968—as well as in the actual pandemic—almost a year passed before the virus made it around the globe. The difference is crucial, because developing and mass-producing a vaccine may take as long as 6 months. Few countries can hope to be spared that long.

Two waves

The toll of the pandemic would depend largely on the attack rate and the mortality rate—two unpredictable factors that can change during an outbreak. Spanish flu, for instance, came in two waves: One, in the spring and summer of 1918, caused widespread disease but few deaths; another, much more vicious wave the following autumn and winter killed half a million people in the United States alone. Presumably, the virus had evolved to become more virulent.

When trying to predict the course of the next pandemic, however, most modelers look more to 1957 and 1968 than to 1918. That's in part because much more is known about the virology and epidemiology of those epidemics, which makes modeling easier. Still, Longini admits that the later pandemics make for rosier outlooks, and the MIDAS group is now collecting data to tackle the 1918 pandemic.

When Meltzer and two CDC colleagues estimated the economic impact of a pandemic on the United States in a 1999 study, they used conservative attack and mortality rates comparable to those in the milder pandemics. Even then, a pandemic could cause between 314,000 and 734,000 hospitalizations and claim between 89,000 and 207,000 lives, they found. Even the lower figures would overwhelm the U.S. health system, says Meltzer: Hospitals were under severe stress when the 1999–2000 flu season was worse than usual.

The team put the economic cost of a 1968-style pandemic for the United States at somewhere between $71.3 billion and $166.5 billion. Using a different set of assumptions, including lower health care costs, Jeroen Medema of Solvay, a vaccine company in the Netherlands, arrived at about $167 billion for all developed countries combined. Both studies, however, included only direct medical costs and lost productivity as a result of disease and

CREDITS: K. BUCKHEIT/*SCIENCE*; SOURCE: REBECCA GRAIS/JOHNS HOPKINS UNIVERSITY

death. A pandemic would almost certainly cause economic disruption that would multiply the cost several-fold. (Asian economies suffered incalculable losses from the SARS outbreak.)

Vaccines would curb the toll, but supplies would be short in the beginning, Meltzer says—as would drugs and attention from doctors and nurses. "Who will get a hospital bed—a 90-year-old grandmother or a 30-year-old mother of two children? People in America are not used to that kind of rationing," Meltzer says, although they're getting a taste of it now that manufacturing problems have abruptly cut the yearly flu vaccine supply in half (see p. 385).

In an as-yet-unpublished paper, Longini and his colleagues show that, when a vaccine is in short supply, different objectives can lead to radically different strategies during relatively mild pandemics. When reducing mortality is the primary goal, for instance, it's best to vaccinate the elderly. When trying to reduce the number of cases or reduce the economic fallout, it would be better to start with schoolchildren.

But so far, there's been little discussion about such priorities and even less consensus. When CDC and other organizations convened a meeting of more than 125 public health experts from 46 states in 2002, participants were asked which of five goals should get top priority during a pandemic: reduce disease, reduce deaths, ensure that essential services continue, limit the economic impact, or ensure "equitable" distribution of scarce resources. None received more than 50% of the votes. "We need a national debate now about these questions," Meltzer says. "When you have a pandemic, it's not a good time to have a discussion with your doctor about the ethics of rationing."

If handled badly, such choices may increase the risk of social upheaval, says Monica Schoch-Spana, a senior fellow at the University of Pittsburgh's Center for Biosecurity. Today's public is likely to become disillusioned when it finds that the government can't offer protection. "There's always the operating assumption that some expert somewhere knows what to do," she says. Clearly explaining the choices as well as the uncertainties is going to be essential, she says.

Retired historian Alfred Crosby, an expert on the 1918 pandemic, is worried about panic, too. But it needn't happen, he notes—the next pandemic may be of the mild rather than cataclysmic variety. Says Crosby: "I wish us all luck." **–MARTIN ENSERINK**

Facing Down Pandemic Flu, the World's Defenses Are Weak

A lack of interest in developing pandemic flu vaccines and a dearth of antiviral drugs have left the world vulnerable to a global outbreak

At a hotel meeting room outside Quebec last March, 35 health officials and others from the world's seven leading industrialized countries and Mexico passed around a vial of bitter-tasting white power. If Asia's potent H5N1 bird flu assumes a form transmitted between humans, this drug, oseltamivir, would be the world's only initial defense against a pandemic that could kill millions of people. But oseltamivir, sold as Tamiflu, is made by only one company, Roche, at a single plant in Switzerland. "We are living in a brave new world where we only have one drug," says flu expert Arnold Monto of the University of Michigan, Ann Arbor, who spoke before the working group meeting of the G7+ Global Health Security Action Group.

That grim assessment is one indicator of the world's vulnerability to pandemic influenza. Most virologists say a pandemic is a virtual certainty within the next few decades, if not from H5N1 then from another avian flu strain (see p. 392). When that happens, public health officials will have two tools to battle the disease: antiviral drugs and vaccines. But although research has produced effective new antivirals, they are expensive, and global supply falls far short of need. And a promising genetically engineered vaccine against H5N1 is still an experimental product only just now being tested in people.

After years of warning from flu experts, governments are finally beginning to respond. Some countries are starting to stockpile antivirals. The United States in August unveiled a draft pandemic flu plan; it is also launching clinical trials of an H5N1 vaccine and will pay Aventis Pasteur $13 million to manufacture 2 million doses. "There's a lot of momentum," says virologist Robert Webster of St. Jude Children's Research Hospital in Memphis, Tennessee.

Priority list. Pandemic vaccines and antivirals will likely have to be rationed to protect the vulnerable, such as children and the elderly.

CREDITS: JOSE LUIS PELAEZ INC./CORBIS

But even that is not enough, say global flu experts. Of the world's 12 major flu vaccine manufacturers, so far only two are willing to tackle the financial, regulatory, and patent issues involved in making a new pandemic vaccine, mainly for the U.S. market. Companies in other countries also need to be developing emergency products, flu experts say. Moreover, only 15 countries have pandemic flu preparedness plans that lay out how scarce vaccines and antivirals will be distributed, notes World Health Organization (WHO) virologist Klaus Stöhr.

As worries intensify, flu experts are exploring a controversial alternative: pooling available supplies of antiviral drugs to stamp out an incipient pandemic in Asia. But whether countries will voluntarily ship their own precious stockpile overseas to fight a faraway plague remains to be seen.

A clear and present danger

The United States last geared up for pandemic flu in 1976, after swine flu broke out in Fort Dix, New Jersey. Within 10 months, the country produced 150 million doses of vaccine and vaccinated 45 million people. But the virus didn't spread, and critics said the government had jumped the gun. That led to the first U.S. pandemic flu plan.

The need to rethink such plans became apparent in 1997, when an outbreak of H5N1 avian flu in Hong Kong killed six people. Unlike previous pandemic strains, H5N1 did not first combine with a human flu virus in

Published by AAAS

pigs; instead it jumped directly to infect humans. This transmissibility and the virus's potency raised the risk that the avian virus could mix with a human flu virus inside a person to yield a deadly pandemic strain. Worries intensified when researchers realized that the tried-and-true method for making flu vaccine in eggs probably would not work with the new avian strain.

Flu vaccines are traditionally made by infecting eggs with a target virus and a nonpathogenic strain that grows well. In the eggs the viruses mix their eight genes. Manufacturers then select a strain with genes for neuraminidase and hemagglutinin (two glycoproteins on the virus's surface) from the target virus, and the rest from the normal flu strain; inactivated virus is then used to make vaccine. But H5N1 kills eggs.

A solution exists: reverse genetics (*Science*, 27 February, p. 1280). Using this technique, the two genes for neuraminidase and hemagglutinin, as well as the six genes from a safe virus, are cloned in bacterial DNA and then reassembled. With highly virulent strains like H5N1, the hemagglutinin gene is first modified to reduce its pathogenicity so the seed virus can be grown in large quantities in eggs. Using reverse genetics, teams at St. Jude and the U.K.'s National Institute for Biological Standards and Control (NIBSC) each produced an attenuated Vietnam H5N1 strain within 3 to 4 weeks earlier this year—"clearly a phenomenal advance," notes Iain Stephenson of the U.K.'s Leicester Royal Infirmary.

Making a candidate vaccine is just the first step; it then has to be tested in humans. Trials of pandemic-like vaccines in the 1970s and since have found that because people have no previous exposure to these viruses' coat proteins, they will likely need two doses plus high levels of antigen. Even then, the vaccine may not work without an adjuvant, a compound that makes the vaccine more immunogenic.

To assess dosage for the reverse-genetics vaccine against Vietnam H5N1, the U.S. National Institute of Allergy and Infectious Diseases (NIAID) expects to begin clinical trials later this year, using lots made by Chiron and Aventis Pasteur from the St. Jude seed strain. Last month, the U.S. Department of Health and Human Services (HHS) also announced that Aventis Pasteur will manufacture antigen for perhaps 2 million doses, depending on how much the clinical trials show is needed. Besides providing a stockpile for health workers exposed to H5N1, "we want to get these manufacturers playing with it" so they can design adequate worker protections and see if the vaccine grows well in eggs, says NIAID's Linda Lambert. The institute also plans to test a vaccine against H9N2, another bird flu strain.

As part of the U.S. draft pandemic flu plan, HHS also disbursed $50 million this year and plans to spend $100 million in 2005 to help ensure that companies have enough eggs year-round. The funds will also support development of an alternative to using eggs—producing vaccine with cell culture using fermenters—an advance that should eventually expand "surge capacity." Under the U.S. plan, "potentially everybody" would get pandemic vaccine, says Bruce Gellin of HHS, although no timeline has been set for reaching this goal.

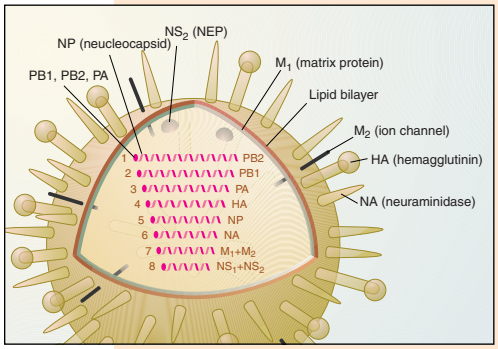
Supply-side economics

So far, only the United States is putting serious money into testing reverse-genetics flu vaccines. And the country is operating "with its own interests in mind," says Stöhr—not to supply the world. (Outside the United States, Japan has plans for trials starting next year, and Aventis Pasteur in France is making test lots of NIBSC's H5N1 seed strain for European trials.)

David Fedson, a retired former medical director for Aventis Pasteur in France, points out that companies in just nine countries in Europe produce 85% of the world's flu vaccine, so if governments decide to impound vaccine to protect their populations (as the

Searching for All-Powerful Flu Weapons

Influenza virus is a shape-shifter, constantly mutating into new pathogenic strains. Every year, companies have to design an entirely new flu vaccine to match the predicted strain's outer coat of proteins. Likewise, to fight a new pandemic strain, researchers would have to start from scratch (see main text), a process that could take 6 months. The Holy Grail of flu research is a vaccine that works against all strains. Many labs and companies are working on this, as well as more effective antiviral drugs. Possible approaches include:



In the crosshairs. Efforts to develop a universal flu vaccine have focused on the virus's conserved proteins.

DNA vaccines. To create a universally protective flu vaccine, researchers are focusing on virus proteins that are conserved among strains or that don't mutate much. A team led by immunologist Suzanne Epstein of the U.S. Food and Drug Administration has shown that a DNA vaccine containing genes for an inner protein, NP, as well as M (matrix) proteins, can work against avian flu. These vaccines deliver strands of DNA into cells, causing the cells to make the antigen themselves. This stimulates various immune responses, including T cells, that provide broader immunity than do vaccines containing only antigen. Live virus also does this, but DNA vaccines are safer and can be produced quickly.

As Epstein's team reported in *Emerging Infectious Diseases* in August 2002, their vaccine, injected into mice, provided partial protection against two strains of H5N1 avian flu. The mice still got sick from the more virulent strain, but half survived a challenge dose that otherwise would have killed them. Such a vaccine could be used to reduce mortality until a matching vaccine became available, Epstein suggests. Others are working on ways to get DNA vaccines to provoke an even stronger immune response, for example by boosting gene expression, using bioengineered proteins, or including additives called adjuvants.

RNA interference. This technique, which involves inserting into cells snippets of RNA that stick to a protein complex that degrades matching viral RNA, could be used as an antiviral to treat flu. In a pair of papers published in the 8 June 2004 issue of the *Proceedings of the National Academy of Sciences*, Jianzhu Chen's team at the Massachusetts Institute of Technology and Epstein's team showed that small interfering RNA constructs with sequences from flu NP and PA genes protected mice against H5 and H7 avian flu subtypes.

New antiviral drugs. To improve on traditional antivirals, molecular biologist Robert Krug's lab at the University of Texas, Austin, is targeting a flu virus protein called NS1 that shuts down the cell's own production of virus-fighting proteins. Because the virus can't avoid using NS1, "we know this is an excellent target," Krug says. A collaborating lab has begun screening for molecules that block NS1 and could be potential drugs.

The problem may be getting companies interested, Krug says. He points to the fate of Relenza (zanamivir), an inhaled drug that may be more impervious to flu virus resistance than oseltamivir, or Tamiflu, the leading flu drug. GlaxoSmithKline cut back its marketing of Relenza in 2000 in response to disappointing sales. **–J.K.**

CREDIT: P. HUEY/*SCIENCE*

United States did during the 1976 swine flu episode), other countries will be in trouble. The United States—which has only one major domestic supplier, Aventis Pasteur in Swiftwater, Pennsylvania—is getting a preview of this scenario this fall, after possible contamination at Chiron's U.K. facility halted use of about 47 million doses of vaccine, half the supply destined for the United States (see p. 385).

Moreover, the world's capacity for making a monovalent pandemic flu vaccine is now 900 million doses, enough for only 15% of the world's population. To stretch the supply, researchers will almost certainly need to use an adjuvant—one that's both cheap and plentiful. Some experts are buzzing about a small trial by GlaxoSmithKline researchers who found that if they used alum to boost an H2N2 vaccine, they needed only 1.875 micrograms of antigen, 12.5% of the normal dose. Alum would also be cheaper than MF59, the adjuvant NIAID plans to test. Adding alum could potentially allow companies to vaccinate 3.5 billion people, or half the world, with two doses of H5N1 vaccine, Fedson says. NIAID isn't pursuing this strategy, however, because no flu vaccine with alum adjuvant has been licensed in United States. "This is a concern," agrees NIBSC's John Wood.

WHO's Stöhr has urged European Commission (EC) leaders to take the initiative in contracting with companies in Europe to test a low-dose pandemic H5N1 vaccine containing alum adjuvant. However, the commission has not yet found the money. "The EC has not the flexibility or the political will," Stöhr says. Companies have little incentive to test pandemic vaccines for a market that may never materialize.

Intellectual-property and liability issues are also major deterrents. The reverse-genetics flu vaccine is licensed by MedImmune, which uses technology from St. Jude. But Mount Sinai School of Medicine and the University of Wisconsin have patents on similar technology. MedImmune has licensed it for research purposes to Aventis Pasteur and Chiron, but if these companies or others wanted to market a vaccine, they would need an agreement with the other patent holders, says Hugh Penfold of the Centre for the Management of IP in Health R&D, a nonprofit in Oxford, U.K. (The U.S. government can assert its patent rights to produce domestic vaccine, but it could not be sold abroad.) Because a reverse-genetics vaccine is considered a genetically modified organism, it would also need special clearance in Europe.

However the vaccine is made, countries would need to pass legislation to shield companies from liability should the vaccine cause serious side effects, as did the swine flu vaccine. Some believe these problems will quickly be solved if a pandemic arrives. "What happens in a crisis is, a lot of the roadblocks get moved," says virologist Maria Zambon of the U.K.'s Health Protection Agency.

Meanwhile, Stöhr notes, countries can get a head start by boosting their capacity to make and deliver regular flu vaccine. Ontario, Canada is a model: Since 2000, the province has offered everyone a free regular flu shot. (Earlier this year, Canada also unveiled a pandemic plan that includes paying one company to manufacture pandemic vaccine for all 32 million Canadians.) Fedson notes that a similar policy in the United States could help guarantee annual flu vaccine supplies and avoid debacles like this year's vaccine shortage, which he hopes will be a "watershed event."

A Rogue's Gallery of Avian Flu Viruses Infecting Humans

STRAIN	LOCATION	INFECTIONS
H5N1	Hong Kong, 1997	18 cases, six deaths
H9N2	Hong Kong, 1999	two cases
H5N1	Hong Kong, 2003	two cases, one death
H7N7	Netherlands, 2003	83 cases of conjunctivitis, one death
H9N2	Hong Kong, 2003	one case
H5N1	Vietnam, 2004	27 cases, 20 deaths
H5N1	Thailand, 2004	16 cases, 11 deaths

CREDITS: IMAGE: C. GOLDSMITH/CDC; SOURCE: ADAPTED FROM TABLE 2 IN I. STEPHENSON *ET AL.*, *THE LANCET INFECTIOUS DISEASES* **4**, 499 (2004) AND WHO LATEST DATA WHICH INCLUDE ONLY LABORATORY-CONFIRMED CASES

Stopgap measures

Even if companies worldwide had the ability and commitment, it could still take 4 to 6 months to manufacture a reverse-genetics vaccine matching a new pandemic flu strain. That leaves antivirals as the first response. Of the two classes of flu antivirals, those that, like Tamiflu, target neuraminidase are considered the best choice because the flu virus is less likely to develop resistance. Roche says that preclinical studies suggest that Tamiflu will be effective against H5N1. Ira Longini, a modeler at Emory University in Atlanta, estimated in a 1 April 2004 paper in the *American Journal of Epidemiology* that a course of antivirals given prophylactically to 80% of the exposed U.S. population for 8 weeks could be as effective as a vaccine in preventing deaths and disease. Although that could require up to 2 billion doses, an unrealistic number, less would be needed if the virus appeared only in some locations.

Some countries, such as Australia, are building sizable stockpiles of Tamiflu. Japan has enough for 20% of its population; the United States can treat 1 million people and hopes to acquire more of the drug. But not all countries can afford Tamiflu, which costs $8 to $10 per course (two pills a day for 5 days) in bulk, Monto notes. And Roche can only make 7 million treatments a year right now (although the company says it can meet all current orders and is expanding capacity).

Most developing countries are in far worse shape. A meeting organized by the Sabin Vaccine Institute in New Canaan, Connecticut, later this month will explore ways to increase vaccine manufacturing capacity in countries such as India. But Africa is "a big, big, question. Without a doubt, the virus will get there. … The situation will be much, much worse than anywhere else. Access to vaccines will not be an option, let alone antivirals," Stöhr says.

With preparations lacking, some experts are mulling whether a mobile stockpile of antivirals could be used to wipe out an incipient pandemic at the source by treating everyone in contact with a patient. This might be feasible, given improvements in WHO surveillance for potential pandemic flu viruses, says Nancy Cox of the U.S. Centers for Disease Control and Prevention. HHS is spending $5.5 million to help countries in Asia begin or improve surveillance for human flu strains, she adds.

Some experts suspect that a pandemic hybrid virus will not be very efficient at human-to-human transmission at first, so it will spread slowly. "We might have a narrow window of opportunity to extinguish it before it becomes a wildfire," says Stöhr.

A consortium of modelers funded by the U.S. National Institutes of Health, including Longini, is looking at the feasibility of stopping a pandemic in Asia if, say, 1 million or 2 million courses of antivirals were available, Cox says. They will present preliminary results at a meeting at Emory in late October. But even if the models suggest it would work, rich countries would need to agree to share their drugs, Stöhr says. The question may be whether an agreement can be reached before the next pandemic arrives.

–JOCELYN KAISER

Vaccinating Birds May Help to Curtail Virus's Spread

As avian influenza continues to ravage Asian poultry, countries are experimenting with a novel control strategy

Fearful that a deadly flu epidemic could be brewing in Asia, some countries are stockpiling drugs, preparing pandemic flu plans, and ratcheting up vaccine production (see p. 394). As these efforts kick into overdrive, animal experts are grappling with the other half of the bird flu equation: the birds. Specifically, they are debating whether a relatively untested strategy of mass vaccination of chickens and other poultry against avian flu will do more harm than good in warding off a human pandemic.

Since its appearance in 1997, global health experts have worried that H5N1 will combine, or reassort, with a human flu virus to produce an easily transmissible strain with H5N1's lethality. To avert such a disaster, last winter and spring seven Asian countries slaughtered more than 100 million birds, decimating the poultry industry. But the virus has resurfaced and appears to be endemic in the region. And the more virus in circulation, the greater the chance of a deadly reassortment.

Animal health officials agree that the best ways to curtail H5N1 are increasing surveillance and improving biosecurity, which includes a host of measures intended to prevent diseases from spreading among flocks and to the public. But now, after years of debate, consensus is building that vaccination of at-risk poultry could also be a critical tool in averting a human pandemic. Indeed, in September, alarmed at the spread of H5N1, the Paris-based World Organization for Animal Health (OIE) and the United Nations Food and Agriculture Organization (FAO) strengthened a previous recommendation encouraging consideration of vaccination in conjunction with other control methods.

But there's a catch, explains Alex Thiermann, a veterinarian at OIE: "If improperly done, vaccination could be dangerous." It could enable the virus to circulate undetected among birds, perhaps spurring its evolution. And no matter how helpful poultry vaccination might be, some countries may decide against it for fear that it would jeopardize their export market.

So far, Hong Kong requires vaccination of all poultry. Thailand forbids it. China and Indonesia are selectively vaccinating in regions where the virus has appeared.

Risks and benefits

The clear benefit of vaccination is its ability to reduce the amount of wild virus in circulation. Although vaccination does not always prevent infection—just disease—it takes a much higher dose of virus to cause infection, and vaccinated birds that do become infected shed far less virus than unvaccinated birds. As an added precaution, animal health experts agree that vaccinated birds that become infected should be culled. "By reducing the amount of virus in the environment, you reduce the possibility of the virus spreading to a new flock, and you reduce the risk to humans," says David Suarez of the U.S. Department of Agriculture's (USDA's) Southeast Poultry Research Laboratory in Athens, Georgia.

Balancing act. Inoculating chickens has its perils but is gaining favor as part of a larger control strategy.

For a country to undertake vaccination safely, it first must ensure the quality and efficacy of the vaccine. It must be targeted to the virus in circulation, properly inactivated, and tested to determine the adequate dosage.

Then there's the problem of distinguishing vaccinated birds from birds infected by the wild virus. If the vaccine is derived from the circulating virus, both infected and vaccinated birds would appear positive in antibody tests. This problem has limited the use of avian flu vaccines in the past because it prevents epidemiologists from tracking the circulating virus. It could also make it hard to prove that flocks are disease-free so exports can resume once the disease is stamped out. (The use of vaccines to control highly pathogenic avian influenza is so new that there are few precedents to follow in resuming trade once an outbreak is contained.)

Long-term experience with an avian vaccine in Mexico has raised other concerns, as reported by Suarez and colleagues in the *Journal of Virology* in August. Farmers in Mexico have been immunizing chickens against a low-pathogenicity H5N2 virus with the same vaccine for 7 years. Over time, the virus has mutated, in a process called antigenic drift. Although the vaccine still prevents clinical disease, it no longer reduces the amount of virus shed by the chickens. Suarez believes that widespread vaccination probably contributed to the virus becoming endemic not only in Mexico but in neighboring Guatemala and El Salvador as well. To avoid this, the virus must be monitored and the vaccine updated periodically.

A shift in favor

Despite these hurdles, sentiment began to shift in favor of adding vaccination to other avian flu control measures several years ago. With the increased scale of modern poultry farms, culling in a buffer zone around an infected flock was killing enormous numbers of healthy birds. Some farmers and animal health officials began arguing that vaccination in a buffer zone, instead of slaughter, might be more humane and cost effective.

In addition, studies done at the USDA lab in Georgia and reported in *Avian Pathology* in 1999 and in *Vaccine* in 2000 showed that a vaccine based on one H5 virus subtype might provide cross-protection against several others. If so, vaccinating with a strain that differs from the circulating strain could solve the problem of differentiating vaccinated-but-uninfected birds from infected birds. More recently, researchers at the Tai Lung Veterinary Laboratory of Hong Kong's Agriculture, Fisheries, and Conservation Department tested a vaccine based on an H5N2 strain against the H5N1 strains

CREDIT: COURTESY OF AGRICULTURE, FISHERIES AND CONSERVATION DEPARTMENT OF HONG KONG

Published by AAAS

Asia Struggles to Keep Humans and Chickens Apart

SONG PHINONG, SUPHANBURI PROVINCE, THAILAND—After having 30,000 chickens culled when H5N1 turned up on a farm 2 kilometers away, Boonchu Taeng-orn got serious about biosecurity. When permitted to restock his farm here in the central lowlands 2 hours north of Bangkok, he followed recommendations of Thailand's Department of Livestock Development to the letter. He strung netting from the shed roofs to the tilapia ponds beneath to keep wild birds out. (Biosecurity experts discourage locating chicken coops near open water, but raising tilapia on bird droppings is key to the economics of chicken farming here.) As few workers as necessary go into the sheds, changing first into work clothes kept at the site, walking through a disinfecting mist, and stepping in pails of disinfectant on the way in. The egg crates are disinfected before use, as are vehicles at the gates to each compound. And Taeng-orn follows the all-in, all-out practice: When he fills a shed with new chicks, he keeps them until egg production drops and then sells the entire batch. Sheds and cages are washed and repaired before the next batch arrives. "The emphasis on cleanliness is definitely good. It is more humane for the animals and safer for the workers," Taeng-orn says.

It is also safer for the world. Infectious disease experts agree that keeping zoonotic diseases like H5N1 and severe acute respiratory syndrome from crossing the species barrier into humans will partly depend on the efforts of millions of farmers like Taeng-orn. A greater challenge is to extend such practices to the numerous households that keep backyard chickens. Alex Thiermann, an official with the World Organization for Animal Health, says that large poultry operations in Asia have biosecurity practices on par with farms in the United States or Europe. But in the backyards, there is "no biosecurity at all."

Risk on wheels. Current methods of transporting live animals facilitate the spread of avian diseases.

A key element of Thailand's push to stamp out H5N1 is to educate small holders and require that even backyard chickens be kept in coops to minimize contact with wild birds and family members. Vietnam, too, has launched an education campaign targeting small chicken operations. But no one expects sudden changes in such an age-old practice.

Hong Kong is taking aim at another entrenched custom: It is considering closing its live animal markets. Currently, buyers pick a live chicken at one of more than 800 live animal shops and have it slaughtered on the spot. K. Y. Yuen, a microbiologist at the University of Hong Kong, favors a central slaughtering facility, both to reduce the chances of exposing the general public to avian influenza and to cut the incidence of other infections. "Other advanced countries adopted central slaughter long ago," he says. The government asked for public comment this summer and is now deciding how to proceed.

–D.N.

that caused outbreaks in Hong Kong in 1997 and 2002. Trevor Ellis, senior veterinary officer at the Tai Lung lab, says the vaccine "protected against clinical disease and produced greater than 1000-fold reduction in virus excretion in birds given heavy virus challenge doses."

More convincing than the lab studies was Hong Kong's experience. Since H5N1 first surfaced there in 1997, the territory has progressively strengthened H5N1 biosecurity measures. Despite these efforts, Hong Kong has repeatedly been hit by H5N1 outbreaks. During an outbreak in December 2002 and January 2003, a number of farms were infected. On three of these farms, chickens in infected sheds were culled, but chickens in other sheds were inoculated with a vaccine based on the H5N2 strain. The virus spread to additional sheds on two of these farms, killing some of the recently vaccinated chickens. But as Ellis and his colleagues reported in the August issue of *Avian Pathology*, 18 days after vaccination, when immunity had developed, there were no new cases of disease among the vaccinated birds; intensive monitoring found no evidence of asymptomatic shedding.

In early 2003, Hong Kong added universal vaccination to its control measures. Unvaccinated "sentinel" chickens are placed within each flock, and there is regular serologic and virologic testing. When H5N1 swept through neighboring China early this year, Hong Kong remained virus-free.

Last winter, both South Korea and Japan identified H5N1 outbreaks quickly enough to contain them with limited culling, still the preferred approach. But where stamping out is impractical or uneconomical, vaccination should be considered, says Joseph Domenech, chief of animal health services for FAO.

Hong Kong's experience is not easily translated to other countries, however. Hong Kong's poultry industry is limited to just 150 farms and a handful of families raising backyard chickens. The territory is small and has an infrastructure capable of fully monitoring the use of vaccines. Hans Wagner, FAO's regional director, says, "It's a substantial challenge to extend these measures to an entire country"—and expensive. The vaccine alone costs about 7 cents per bird, not counting the labor of injecting or the monitoring that should accompany it. By contrast, FAO consultants and others who have visited China and Indonesia—which are both vaccinating in areas where H5N1 has been reported—noted several shortcomings. Several of the vaccines in use in both countries are based on the H5N1 strain itself, making it difficult to track the disease. And the use of unvaccinated sentinels and the serological and virological monitoring is spotty at best.

In Thailand, which has reported more than 250 outbreaks in 45 of the country's 76 provinces in the last 3 months, authorities have rejected vaccination, at least for the moment. Yukol Limlamthong, director-general of Thailand's Department of Livestock Development, says they are worried that vaccination might enable the virus to circulate silently among vaccinated birds, exposing farm hands and families to infection. "We don't want to put them at risk," he says. But flu experts elsewhere suspect that commercial concerns factored heavily in the decision.

The OIE Terrestrial Animal Health Code, which governs international trade in animals and animal products, says a country can be considered free of avian influenza if specified levels of surveillance do not turn up the virus—regardless of whether it is vaccinating. But the code is vague and places the burden of proof on the exporting country. Johan Reyniers, a press spokesperson for the European Commission in Brussels, says, "It would ultimately be up to Thai authorities to demonstrate that vaccination is properly implemented."

For now, Thai officials believe it will be easier to convince trading partners that its poultry products are safe if the country can control the disease without vaccination. But whether it can remains to be seen.

–DENNIS NORMILE

With reporting by Xiong Lei in Beijing.

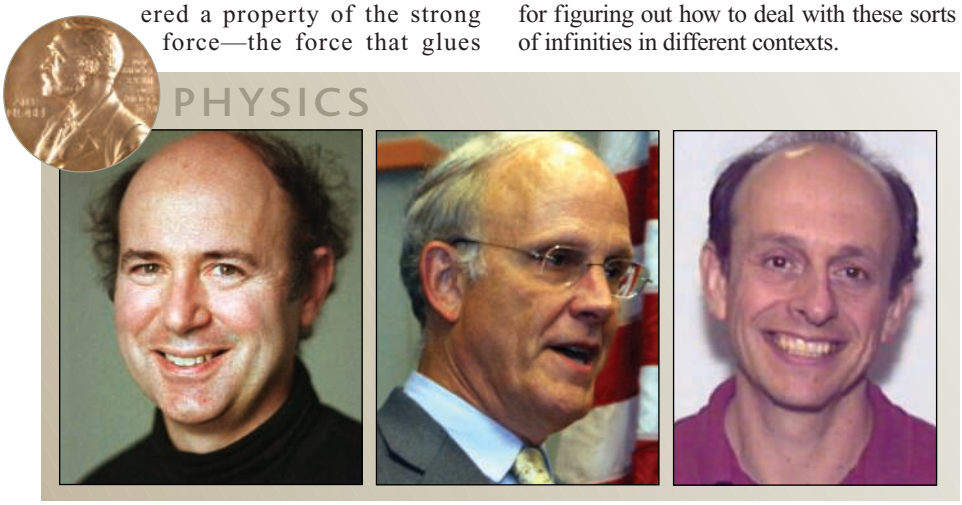
CREDIT: HANS WAGNER/U.N. FOOD AND AGRICULTURE ORGANIZATION

Published by AAAS

Laurels to Three Who Tamed Equations of Quark Theory

It might be fun to blow things up, but this year's winners of the Nobel Prize in physics earned the plaudits of their colleagues with a discovery that does the opposite: It prevents equations that describe one of the fundamental forces of nature from exploding.

The three new laureates, Frank Wilczek, David Gross, and H. David Politzer, discovered a property of the strong force—the force that glues quarks to one another—known as "asymptotic freedom." Not only did the idea explain some baffling experimental results in particle colliders, but it also showed how to keep the equations that describe the strong force from producing troublesome infinities. "They made the discovery and saw the significance of it," says Niels Kjaer Nielsen, a physicist at the University of Southern Denmark in Odense. "[The prize] is fully deserved."



Exorcists. Frank Wilczek (*left*), David Gross, and H. David Politzer banished unwanted infinities.

Particle physics is swimming with infinities: places where the equations that describe the behavior of a particle seem to blow up into a meaningless jumble of singularities. One reason is that every region of space, even the deepest vacuum, is seething with "virtual" particles that pop in and out of existence—and these particles make even the simplest concepts very difficult.

For example, an electron is surrounded by a cloud of evanescent particles. When scientists try to gauge its charge, the cloud "screens" the naked electron and hides some of the charge from view. If you could somehow worm a measuring instrument through the cloud, getting closer and closer to the bare electron at the center, you would see the measured charge get greater and greater as you penetrate the screen of virtual particles. Strictly speaking, the plain-vanilla equations of the Standard Model of particle physics say that the charge increases without bound at smaller and smaller distances. In other words, the equations blow up. Scientists have come up with mathematical coping mechanisms to get around this problem; the 1965 and 1999 physics Nobels were given for figuring out how to deal with these sorts of infinities in different contexts.

In the early 1970s, physicists studying the strong force were beating their heads against a similar problem. But the infinity-coping techniques developed for the electric force (and for the weak force, which is responsible for phenomena such as nuclear decay) didn't work for the strong force—until Wilczek, Gross, and Politzer made a counterintuitive discovery.

In 1973, Politzer, currently at the California Institute of Technology in Pasadena, and, separately, Wilczek, at the Massachusetts Institute of Technology, with Gross, at the Kavli Institute for Theoretical Physics in Santa Barbara, California, realized that, unlike the electric (and weak) forces, the strong force gets weaker at close range—much as a taut spring relaxes when the ends are brought close together. As a result, virtual particles "screen" quarks in a very different way from how they screen electrons: The virtual particles—gluons—that surround and interact with a quark feel one another's presence in a way that the virtual particles that surround and interact with an electron—photons—don't. Stick a particle right next to a quark, and it wouldn't feel the strong force at all; it would be "asymptotically free" from the strong force, and quarks forced into close proximity would behave more or less like hard little unbound particles rather than a sticky clump. That is precisely what experimentalists at the Stanford Linear Accelerator Center in California had found a few years earlier by scattering electrons off protons. Turned around, asymptotic freedom explains why quarks are never found roaming free from one another: At large distances and low energies, the strong force is too powerful to overcome.

Particle theorists have long anticipated this award, and Wilczek was no exception. "I'd be lying if I said it was unexpected," he said with a laugh. "I thought it was an important theory, and the data in favor of it has been clear for at least 20 years." And in that time, thanks in part to this year's laureates, our understanding of the fundamental constituents of forces and particles has exploded.

–CHARLES SEIFE

Gold Medal From Cellular Trash

The cell's trash collectors, which control an internal system of protein disposal, are celebrated in this year's Nobel Prize in chemistry. The discoverers of this system, Aaron Ciechanover and Avram Hershko of the Rappaport Institute at the Technion–Israel Institute of Technology in Haifa and Irwin Rose of the University of California, Irvine, share the prize for work that established how a protein called ubiquitin, with several helpers, tags and delivers other proteins for recycling. The prizewinning experiments were "an extraordinary tour de force of classical biochemistry," says Kim Nasmyth of the Research Institute of Molecular Pathology in Vienna, who helped clarify the role of ubiquitin in cell division.

While most biochemists in the 1970s were studying how cells make proteins, Hershko and Rose became interested in a less-studied puzzle: why a cell requires energy to break down proteins. In 1979, Hershko and Ciechanover, then a graduate student, pursued this topic with a series of experiments while on sabbatical at Rose's lab at the Fox Chase Cancer Center in Philadelphia. The result was a pair of papers published in 1980 in the *Proceedings of the National Academy of Sciences* revealing that proteins destined for destruction were covalently bonded—a process that requires energy—to a small protein the team called APF-1. That

CREDITS (LEFT TO RIGHT): COURTESY OF MIT; COURTESY OF UCSB; COURTESY OF CALTECH

Published by AAAS

CHEMISTRY

Cleaning up. Avram Hershko (*left*), Aaron Ciechanover, and Irwin Rose unraveled ubiquitin's role.

protein later turned out to be ubiquitin, which had been identified by other researchers a few years earlier, and which is found in eukaryotic organisms from yeast to mammals—hence its name.

The biochemists went on to show that three additional enzyme families, called E1, E2, and E3, work together to attach ubiquitin to proteins destined for disassembly. They and others subsequently showed that ubiquitin then delivers the doomed proteins to the proteasome, a large complex that breaks down the chemical bonds holding proteins together and releases their amino acid building blocks for reuse. Ciechanover says the discoveries honored by the Nobel committee helped explain how the protein-degrading proteasome can coexist with proteins in the cell's cytoplasm without breaking down the wrong ones. "The shark and the bait are living together peacefully, and they will interact only following the tag from ubiquitin," he says.

A decade after the trio made their discoveries, researchers began to realize that ubiquitin's job was more than simple trash collecting. The protein and its enzyme helpers play a role in the cell's proofreading of newly minted proteins, targeting faulty ones for destruction. The ubiquitin system also helps regulate cell division, where it controls the swift buildup and breakdown of proteins that drive the cell cycle. It plays a crucial role in triggering DNA repair and apoptosis by influencing cellular levels of the tumor suppressor protein p53. And it helps regulate the signaling protein NF-κB, which triggers immune and inflammatory responses.

In recent years researchers have begun to piece together even more exotic roles for ubiquitin, including helping to transport proteins from the cell surface to the interior (*Science*, 13 September 2002, p. 1792). On the negative side, the protein is involved in enabling viruses such as HIV and Ebola to make their way to the cell surface after replicating inside the cell.

Drug companies also think they may find a way to exploit ubiquitin and its helpers. By blocking the system, researchers have been able to halt cell division in cancerous cells. One drug that blocks the action of the proteasome was recently approved for treating patients with multiple myeloma, a type of leukemia.

Nasmyth says the new Nobel laureates had no way of knowing how important their find would be. "This is a discovery that has impacted every single branch of biology and is a beautiful bit of chemistry," he says.

–GRETCHEN VOGEL

Medicine, Peace Prizes

For details of the 2004 Nobel Prize in physiology or medicine, awarded to Richard Axel and Linda Buck for their work on olfaction, see *Science*, 8 October, p. 207. Coverage of the Nobel Peace Prize can be found on page 391 of this issue.

Macroeconomists Showed Why Good Intentions Go Wrong

It's no great insight to realize that governments behave in a less-than-optimal manner. Understanding why—that's another story. This year's Bank of Sweden Prize, otherwise known as the Economics Nobel, goes to Finn Kydland and Edward Prescott, two economists who figured out why good governments do bad things to good people. "I'm still high. It's a great event," says Robert Lucas, an economist at the University of Chicago, who won the prize in 1995. "These are great economists."

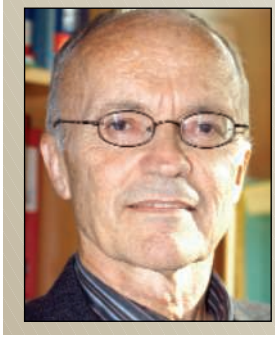
In the mid-1930s, economist John Maynard Keynes came up with a successful framework for analyzing broad trends in unemployment, consumption, production, and inflation. The Keynesian picture seemed to promise a utopia, a way to keep inflation and unemployment in check through an optimal strategy of setting taxes and interest rates and other tools of economic policy. But as with all utopias, an ideal economic policy turned out to be a pipe dream. Inflation and unemployment often fluctuated out of control, and occasionally a government's well-intentioned actions would make matters worse. Sometimes, the seemingly impossible would happen. For example, in the late 1970s, inflation and unemployment rose dramatically at the same time—something that the Keynesian picture forbids.

In the late 1970s and early 1980s, Prescott, of Arizona State University in Tempe, and Kydland, of Carnegie Mellon University in Pittsburgh, Pennsylvania, and the University of California, Santa Barbara, figured out why optimal-seeming fiscal strategies sometimes have suboptimal results. The two showed that governments have trouble committing to a policy; this lack of commitment leads to a credibility problem, which, in turn, can lead to an undesirable outcome. "The effect of a tax cut today depends on whether people think it is permanent or just temporary," says Lucas. Inserting that insight into the mathematical models of macroeconomics changed the way economists think, he says: "It was a huge break from what all of us were doing at the time."

"It just hit us in the nose," says Prescott. The new approach also led to a better understanding of the causes of business cycles that rattle through an otherwise stable economy. As Prescott and Kydland discovered, it's in the equations: The best-laid schemes o' mice an' men gang aft agley.

–CHARLES SEIFE

ECONOMICS



Snafu experts. Finn Kydland (*left*) and Edward Prescott modeled how short-term lapses undermine economic policy.

CREDITS (CLOCKWISE FROM TOP LEFT): ASSOCIATED PRESS; COURTESY OF UCI; AFP/GETTY IMAGES; AFP/GETTY IMAGES

Published by AAAS

Ongoing Controversy over Romanov Remains

THE FIELD OF ANCIENT DNA ANALYSIS HAS faced numerous obstacles and setbacks in its path to legitimacy. Yet another setback was showcased in the news story "Buried, recovered, lost again? The Romanovs may never rest" (R. Stone, News Focus, 6 Feb., p. 753). Much was made of a report by A. Knight *et al*. (*1*) that claimed to be a failed attempt to "replicate the findings" of a previous DNA analysis of the putative remains of Tsar Nicholas II of Russia, the Empress Alexandra, and three of their daughters (*2*).

Tsar Nicholas II of Russia, his wife Empress Alexandra, and their five children, circa 1907.

CREDIT: UNDERWOOD & UNDERWOOD/CORBIS

Knight *et al*. did not, in fact, test the skeletal material in question, but used a new maternal reference sample for Alexandra: an 86-year-old finger putatively from Alexandra's sister, Grand Duchess Elisabeth of Russia. We cannot see why anyone would consider this a superior DNA source to the modern-day blood sample from Alexandra's grandnephew Prince Phillip of Great Britain that was analyzed previously. Moreover, the finger showed a mixture of mitochondrial DNA (mtDNA) sequences from different individuals, and in two of four amplifications showed a minority sequence that matched a rare sequence motif shared by Prince Phillip and Alexandra. The results of Knight *et al*. end in a fizzle. The fuss has been caused by their claim that recent developments in the ancient DNA field (*3*, *4*) constitute "certain evidence" of the fallacy of the Gill *et al*. (*2*) testing, because of amplicon sizes involved.

This bald assertion naively elevates a well-established truism of ancient DNA—that it is fragmented in length—to categorical law, ignoring the breadth of ancient/forensic DNA literature and experience and the range of considerations that are part of determining ancient DNA authenticity. Knight *et al*. repeatedly assert that DNA fragments greater than 250 base pairs (bp) do not exist in samples as little as 70 years old. However, DNA preservation depends on both the age of a sample and the environmental context, with comparatively cold temperatures greatly favoring DNA preservation (*5*, *6*). Instances of remarkable preservation include recovery of a 1.7-kb fragment of a single-copy gene from a 156-year-old dried specimen (*7*), 1.6 kb from 560-year-old avian bones (*8*), and 438 bp from a 3350-year-old moa bone (*9*). Successful amplification of 522 bp of mtDNA from a 20,000-year-old ground sloth coprolite from Utah (*10*) suggests that amplification of DNA fragments twice that length from bones 300 times younger is far from implausible. Moreover, Knight *et al*. fail to acknowledge that the 1223-bp amplicons of Gill *et al*. (*2*) were used only in the first round of a 62-cycle nested PCR protocol.

We explored the use of nested and non-nested PCR on six degraded skeletal extracts of known authentic sequence: ~60-year-old bones of three individuals recovered from a crash in temperate coastal Alaska (from a lower latitude than the Romanov remains) and three from temperate Asia (~50 years old). Using standard single round PCR, we did, in fact, obtain successful PCR and authentic sequence with 1200-bp amplicons for two of the three Alaska bones, but not with any others. However, when a nested protocol similar to Gill *et al*. (*2*) was employed using 1200-bp primers in the first round and 221-bp primers in the second, we obtained authentic sequence from all three Alaskan bones and two of the Asian bones. These results suggest there is nothing implausible about the results of Gill *et al*. (*2*).

Published by AAAS

Knight *et al.* fail to cite the replication of mtDNA results of the Tsar in an independent laboratory with different methods (*11*), an important criterion of ancient DNA authenticity. Furthermore, another criterion of ancient DNA authenticity in the Romanovs is also met: The results make sense in the genetic context of the investigation. The nuclear short tandem repeats (STRs) are consistent with a mother, father, and three daughters, and there is an mtDNA link of the mother to Prince Phillip, an mtDNA link of the father to living relatives, and shared heteroplasmy with the Tsar's brother (*11*). The chances that these results are from contamination are astronomically slim.

As no reasonable alternate explanation for the data is apparent, or has been offered, we conclude that there is no scientific reason to refute the identification of the Romanovs. Although ancient DNA research will always remain prone to artifacts because of contamination, requiring carefully conducted studies, it should not be put out of the realms of science into some mystic sphere where generalized criteria suggested in review articles are used as dogma to refute otherwise indisputable scientific results.

MICHAEL HOFREITER,[1] ODILE LOREILLE,[2]
DEBORAH FERRIOLA,[2] THOMAS J. PARSONS[2]

[1]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. [2]Armed Forces DNA Identification Laboratory, 1413 Research Boulevard, Rockville, MD 20850, USA.

References
1. A. Knight *et al.*, *Ann. Hum. Biol*. **31**, 129 (2004).
2. P. Gill *et al.*, *Nature Genet*. **6**, 130 (1994).
3. M. Hofreiter *et al.*, *Nature Rev. Genet*. **2**, 353 (2001).
4. A. Cooper, H. Poinar, *Science* **289**, 1139 (2000).
5. E. Willerslev *et al.*, *Curr. Biol*. **14**, R9 (2004).
6. C. I. Smith *et al.*, *J. Hum. Evol*. **45**, 203 (2003).
7. D. M. Hunt *et al.*, *Science* **267**, 984 (1995).
8. D. M. Lambert *et al.*, *Science* **295**, 2270 (2002).
9. A. Cooper *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8741 (1992).
10. H. Poinar *et al.*, *Curr. Biol*. **13**, 1150 (2003).
11. P. Ivanov *et al.*, *Nature Genet*. **12**, 417 (1996).

RICHARD STONE'S NEWS STORY "BURIED, recovered, lost again? The Romanovs may never rest" (News Focus, 6 Feb., p. 753) highlights a study by A. Knight and colleagues in which they analyzed a shriveled finger said to be from Grand Duchess Elisabeth, a sister of Empress Alexandra of Russia (*1*). They recovered a mitochondrial DNA (mtDNA) sequence of unknown origin from the finger and concluded that the previous identification of the remains found at Ekaterinburg, Russia, as the Romanovs (*2*) was inconclusive.

The arguments of Knight *et al.* are illogical. The claim that they identified the correct mtDNA sequence is not substantiated, and their anecdotal evidence of the origin of the finger is irrelevant to this DNA evidence. Their reported mtDNA sequence did not match that previously obtained from remains formally identified as those of Alexandra and three of her daughters, and from blood from Prince Philip, the Duke of Edinburgh, a known grandnephew of Alexandra (*2*). They also criticize the original investigation of the purported Romanov remains by physical anthropologists.

In our investigation (*2*), we evaluated the DNA evidence using a Bayesian approach (*3*). The prior odds for the non-DNA anthropological and historical evidence were obtained from a relevant expert, and we presented the DNA data in an objective probabilistic framework to allow others to reach a conclusion based on their interpretation of the DNA and non-DNA evidence. The Russian authorities accepted that the remains were those of the Romanovs after considering all the expert evidence.

Knight *et al.* assert that our findings were the result of contamination. Although contamination is a potential problem in the analysis of biological samples containing small amounts of DNA, such as old bones, our respective laboratories established a number of principles governing this type of work in forensic identification and ancient DNA research well before the Romanov investigation (*4*). In particular, Knight *et al.* failed to note that the DNA extractions and mtDNA sequencing of samples of the nine Ekaterinburg skeletons were replicated blindly by one of us (E.H.) in a separate laboratory (*2*). A key finding, the characterization of mitochondrial heteroplasmy in the putative Tsar's remains, was also replicated independently (*5*). The allegation that bone samples were contaminated by present-day DNA of a maternally related individual is untenable, as we approached the relatives after we had typed the bones. In addition to comparing mtDNA of the putative Romanovs with that of living relatives, the presence of a family group among the nine bodies was confirmed by STR

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

analysis. The sexing of skeletons by physical anthropologists was confirmed by analysis of the amelogenin gene. Importantly, rather than resulting in incorrect inclusions, random contamination generates inexplicable DNA profiles that lead to exclusions (*6*, *7*).

Knight *et al.* used cloning to prove that the mtDNA sequence from the Elisabeth relic was genuine, while asserting that our results were flawed. Cloning does not guarantee that the product is not contamination, because contaminating DNA can be cloned as readily as authentic bone DNA. However, we did clone the mtDNA amplification products to resolve the issue of the heteroplasmy in Tsar Nicholas, although the remaining samples gave reproducible results without cloning.

The most logical explanation of the results by Knight *et al.* is that the shriveled finger was not from Elisabeth or that the DNA sequence they recovered was the result of contamination. Their cloning results cannot refute these arguments. Conversely, contamination cannot explain the agreement between the mtDNA sequences of the presumptive Romanovs analyzed independently in three laboratories, or their match with DNA of known living relatives.

PETER GILL[1] AND ERIKA HAGELBERG[2]

[1]Forensic Science Service, 2960 Solihull Parkway, Trident Court, Solihull B37 7YN, UK. E-mail: dnapgill@compuserve.com. [2]Department of Biology, University of Oslo, Post Office Box 1050 Blindern, Oslo 0316, Norway. E-mail: erika.hagelberg@bio.uio.no

References
1. A. Knight *et al.*, *Ann. Hum. Biol.* **31**, 129 (2004).
2. P. Gill *et al.*, *Nature Genet.* **6**, 130 (1994).
3. I. W. Evett, B. S. Weir, *Interpreting DNA Evidence* (Sinauer Associates, Sunderland, MA, 1998), pp.17–21.
4. E. Hagelberg *et al.*, *Nature* **352**, 427 (1991).
5. P. Ivanov *et al.*, *Nature Genet.* **12**, 417 (1996).
6. P. Gill *et al.*, *Forensic Sci. Int.* **112**, 17 (2000).
7. P. Gill, A. Kirkham, *J. Forensic Sci.*, in press.

Response

THE CONCLUSION THAT THE EKATERINBURG remains were those of the Romanovs faced caveats from the forensic perspective (*1*) that have not been acknowledged by the authors of the DNA analyses. They did not respond to requests to provide the "raw" DNA data and for documents of chain of custody. Therefore, we, with the Russian Expert Commission Abroad, conducted an additional DNA investigation (*2*). As we explicitly stated, our main conclusion was based on the reported claim that the authors had obtained sequence "comparable to that produced from the fresh blood" from polymerase chain reaction (PCR) products of 1223 base pairs (bp) in length from each of nine bones (*3*). Generally, published results indicate that only fragments shorter than about 250 bp are obtainable from old tissues not stored in favorable environmental conditions. An independent test of the Ekaterinburg remains, carried out on teeth, was consistent with established molecular behaviors of such samples in that only very short PCR products were obtainable and sequence was of poor quality (*4*). Gill and Hagelberg have not addressed this central issue. Hofreiter and Parsons provide only two examples of results of similar length. One is of tissues of penguins frozen in Antarctica, and the other of carefully preserved tissues of the eye of John Dalton. Likewise, preservation of the avian bones and sloth coprolite was excellent. None of these preservation environments remotely resembles the wet soil of Ekaterinburg, where climate is continental with hot summers. Gill and Hagelberg refer to "an objective probabilistic framework." The prior probability is exceedingly low that nine badly decomposed bones, submerged in wet soil for several decades, can produce PCR products of 1223 bp in length for every tested individual. There are no other such published results. Generally, such results indicate contamination (*5*, *6*).

One of us had suggested to Parsons that studies of bones of similar age and condition, subjected to the methods in (*3*), would be necessary to establish that such results were possible. Now their team has carried out experiments on bones from Alaska and Asia, a first step toward that goal. They do not provide information on results of experiments that duplicate the nested PCR method in (*3*), using the same PCR primers with nested products of about 400 bp in length. Instead, they obtained sequence from a 221-bp product, well within the range of degraded DNA. Sample preservation and their experimental methods and results are not published or revealed in full, and successful PCR of 1200 bp indicates excellent preservation of those two Alaska individuals. Nothing has been accomplished to indicate that the results in (*3*) are plausible. To the contrary, only a 221-bp amplicon could be produced (possibly from endogenous degraded DNA template), but not a 400-bp nested product. This result further supports our conclusion that the results in (*3*) are not plausible.

Gill and Hagelberg imply both degradation and possible misidentification of the Elisabeth sample. Elisabeth's body was identified by those who knew her and placed in a sealed coffin with her name inscribed and kept in a locked crypt. The finger included dried flesh as well as bone,

indicating stable conditions of preservation. Tests of molecular behavior of the finger (*2*) were consistent with an old sample.

There are many shallow mass graves in the Ekaterinburg region, including entire families that resemble the remains in question (*7*–*10*). The grave had been opened many times over the decades with many bones removed and added (*1*, *8*–*10*). The "discoverers" of the grave, Ryabov and Avdonin, removed three skulls in 1979, over a decade before the time of discovery reported by Gill *et al.* (*3*), and took two of them to Moscow (*1*, *6*–*8*). It is documented in a medical report dated 1891 and signed by three Russian naval physicians that the skull bones of the Tsarevich Nicholas had a deep scar from a sword wound (*11*), and there was no trace of this gash in the skull from Ekaterinburg. For purposes of facial reconstruction, crucial reference points were missing from the damaged and decomposed skulls (*8*). Arm and leg bones had sections removed, making it impossible to estimate individual height (*8*). Expert forensic physical anthropologists, including William Maples, strongly objected to the methodology and conclusions (*1*, *8*, *9*).

Our critics confuse repetition with replication. They analyzed bones provided by Russian geneticist P. Ivanov, who had access to all the samples, conducted tests, prepared a report to the Russian government, and then voted on acceptance of that report (*1*). Our test of Elisabeth was replication. We did not cite the tests later conducted in the United States (*12*) because they could have been contaminated from the same source, and the fragment lengths tested were much shorter. Heteroplasmy was not found in a sample from the Tsar's nephew (*13*) and apparently was not found in the Tsar's blood-soaked bandage (*8*). As unlikely as it is to have obtained such perfect mtDNA results, the STR results are even more unlikely without the presence of "fresh" DNA. Gill has stated, "they are probably the oldest samples from which this kind of DNA ever has been extracted" [(*9*), p. 104].

DNA testing by proponents of Romanov identity has been shrouded in secrecy. The possibility of a mismatch between the mtDNA of Prince Philip and that of his sister has been suggested (*8*, *14*). Also, the mtDNA of another relative, Princess Feodora, was found to have a C at position 16111, whereas her mother Princess Charlotte had a T at that position (*15*). All these individuals are expected to carry the mtDNA lineage of Queen Victoria, grandmother of Empress Alexandra and Grand Duchess Elisabeth. There are over 50 living carriers of that lineage. Truly independent tests of some of these individuals, with full disclosure of chain of custody, are now necessary to establish this haplotype. Given present knowledge and inconsistencies, the Ekaterinburg remains cannot be regarded as those of Nicholas II and his family.

ALEC KNIGHT,[1] LEV A. ZHIVOTOVSKY,[2] DAVID H. KASS,[3] DARYL E. LITWIN,[4] LANCE D. GREEN,[5] P. SCOTT WHITE[5]

[1]Department of Anthropological Sciences, Stanford University, Stanford, CA 94305, USA. [2]Vavilov Institute of General Genetics, Russian Academy of Sciences, 117809 Moscow, Russia. [3]Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197, USA. [4]Post Office Box 19754, Stanford, CA 94309, USA. [5]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

References

1. L. A. Zhivotovsky, *Ann. Hum. Biol.* **26**, 569 (1999).
2. A. Knight *et al.*, *Ann. Hum. Biol.* **31**, 129 (2004).
3. P. Gill *et al.*, *Nature Genet.* **6**, 130 (1994).
4. C. Ginther, personal communication.
5. A. Cooper, H. Poinar, *Science* **289**, 1139 (2000).
6. M. Hofrieter *et al.*, *Nature Rev. Genet.* **2**, 353 (2001).
7. A. Summers, T. Mangold, *The File on The Tsar* (Gollancz, London, 1976).
8. M. Gray, *Blood Relative* (Gollancz, London, 1998).
9. R. K. Massie, *The Romanovs: The Final Chapter* (Random House, New York, 1995).
10. E. L. Magerovsky, *Trans. Assoc. Russian-Am. Scholars* **28**, 449 (1996–1997).
11. State Archives of the Russian Federation, Folder 77, Reg. 1, Doc. 701, leaves 12-13.
12. P. L. Ivanov *et al.*, *Nature Genet.* **12**, 417 (1996).
13. E. I. Rogaev, I. V. Ovchinnikov, P. Dzhorzh-Khislop, E. A. Rogaeva, *Genetika* **32**, 1690 (1996).
14. M. Kirk, in Proceedings of International Scientific Conference on "Tsar Case and Ekaterinburg Remains," St. Petersburg, Russia, 26 to 27 April 1998 (in Russian), available at www.tzar.orthodoxy.ru/ost/mnk/8.htm.
15. J. C. G. Röhl, M. Warren, D. Hunt, *Purple Secret: Genes, "Madness" and the Royal Houses of Europe* (Bantam, London, 1998).

Producing Neuronal Energy

I WISH TO COMMENT ON THE REPORT "Neural activity triggers neuronal oxidative metabolism followed by astrocytic glycolysis" by K. A. Kasischke *et al.* (2 July, p. 99) on energy metabolism involving interactions between neurons and astrocytes, and on the commentary to this paper, "Let there be (NADH) light" (L. Pellerin, P. J. Magistretti, Perspectives, 2 July, p. 50). Kasischke *et al.* present evidence for the coupling between oxidative metabolism in the dendritic shaft and glycolytic activity in the astrocyte to provide sustained neuronal energy in the form of adenosine triphosphate (ATP). The ATP produced by mitochondria in the dendritic shaft is presumed to travel into the dendritic spine, the site of the synapse, because there are no mitochondria in the spine. However, 7 years ago (*1*), another energy-producing system, forming ATP, was found to exist in the spine

Published by AAAS

itself, produced by glycolytic enzymes localized in a structure called the postsynaptic density (PSD) at the postsynaptic membrane of the synapse.

The significance of these findings is that the ATP generated at this important synapse site can be used at ion channels there, by various protein kinases there to phosphorylate proteins for signal transduction, and also for protein synthesis in the spine. Thus, there could exist two energy-producing systems, one existing in the astrocyte and in the dendritic shaft in the production of energy for general "housekeeping" metabolic functions, and the other at the postsynaptic PSD site for the processing of nerve signal transduction. With regards to the latter, the communication requirement of the central nervous system almost invites a specialized, localized, structural site for its most important mission.

PHILIP SIEKEVITZ

Rockefeller University, 1230 York Avenue, New York, NY 10021, USA.

Reference

1. K. Wu, C. Aoki, A. Elste, A. Rogalski-Wilk, P. Siekevitz, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13273 (1997).

Response

SIEKEVITZ RAISES AN INTERESTING POINT, and we were well aware of his intriguing proposal of glycolytic generation of ATP in the dendritic spine during synaptic activity. It appears that according to his model, the temporal pattern of the presumed glycolytic response in the dendritic spine would have to closely, if not directly, follow the presynaptic input in order to meet metabolic needs induced by nerve signal conduction. In our experiments, we did not observe a rapid glycolytic burst. In contrast, the glycolytic response did not start until 10 s after stimulus onset and reached its peak after the cessation of the stimulus. However, our described experimental setup was not optimized to detect or locate a presumed brief transient glycolytic burst in the spine.

We have shown that the activity-dependent glycolytic and oxidative metabolic responses in the central nervous system (CNS) are highly compartmentalized between astrocytes and neurons. It might indeed be intriguing to learn whether the (sub)cellular compartmentalization and specialization of the CNS is even more refined on a higher temporal and smaller spatial scale.

KARL A. KASISCHKE,* WATT W. WEBB

School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14853, USA.

*To whom correspondence should be addressed. E-mail: kak32@cornell.edu

CORRECTIONS AND CLARIFICATIONS

Reports: "Decamethyldizincocene, a stable compound of Zn(I) with a Zn–Zn bond" by I. Resa *et al.* (20 Aug., p. 1136). In Reference 13, space group $P\bar{1}$ should be space group $P\bar{1}$.

Special Section on Testing Human Limits: News: "Peering under the hood of Africa's runners" by C. Holden (30 July, p. 637). Kip Keino set an Olympic record at the 1968 Summer Olympics, not a world record. Also, the image on page 638 shows Carl Lewis winning the 4 × 100-meter relay at the 1992 Summer Olympics, not the 400-meter race as stated in the caption. Lewis did not compete in the 400-meter event.

Reports: "Self-assembled hexa-*peri*-hexabenzocoronene graphitic nanotube" by J. P. Hill *et al.* (4 June, p. 1481). On page 1481, in the ninth line of the abstract, "an electrical conductivity" should have been "an electrical resistance." On page 1483, in the 28th and 29th lines of text in the first column, the word "resistivity" should read "resistance."

TECHNICAL COMMENT ABSTRACTS

COMMENT ON "Force-Clamp Spectroscopy Monitors the Folding Trajectory of a Single Protein"

T. R. Sosnick

In a recent report on atomic force microscopy (AFM)–monitored protein folding, Fernandez and Li (Reports, 12 March 2004, p. 1674) concluded that the folding of the single-domain protein ubiquitin does not correspond to transitions between discrete states. The results are inconsistent with solution studies of ubiquitin folding and probably are due in part to chain-tangling in the tethered polyprotein construct used in the AFM studies.

Full text at www.sciencemag.org/cgi/content/full/306/5695/411b

RESPONSE TO COMMENT ON "Force-Clamp Spectroscopy Monitors the Folding Trajectory of a Single Protein"

Julio M. Fernandez, Hongbin Li, Jasna Brujic

The mechanical stretching of ubiquitin chains creates a novel and well-defined starting point for observing folding trajectories. These initial conditions, which are never reached through chemical denaturation, add new physics to the folding problem. In contrast to chemical denaturation, mechanical unfolding and refolding of tandem modular proteins occur in vivo. It is therefore not likely that such events involve chain-tangling.

Full text at www.sciencemag.org/cgi/content/full/306/5695/411c

Comment on "Force-Clamp Spectroscopy Monitors the Folding Trajectory of a Single Protein"

Taking advantage of major improvements in their atomic force microscopy (AFM) apparatus, Fernandez and Li were able to follow single-molecule refolding trajectories of the ubiquitin protein (*1*). They observed a rich variety of kinetic behavior. Using a polycistronic version of ubiquitin, lengths of three to eight tethered proteins were picked up at random locations and unfolded using a pulling force. Upon relaxation of the force, refolding occurred in continuous stages. The results were interpreted in terms of a folding scenario with no defined kinetic barrier between the unfolded and folded states.

Monomeric ubiquitin free in solution has been demonstrated to fold in a barrier-limited process (*2*–*5*), often in a two-state manner (*6*–*15*) without the multiple early collapse phases (*12*) seen in the AFM studies. Two-state behavior persists even when there is transition-state heterogeneity (*11*). The discrepancy between the ensemble and AFM measurements cannot be solely attributed to the measurement of single molecules; other single-molecule measurements, in which the proteins were monomeric and free in solution, were fully consistent with analogous solution results that show two-state folding and discrete transitions (*16*, *17*).

One suspects that the nondiscrete folding behavior observed for tethered proteins in the AFM studies was due to the intimacy of the multiple ubiquitin chains. In free solution, detectable aggregation of refolding ubiquitin occurs at 2 μM concentration (*15*), which is resolved on the millisecond-to-second time scale (*3*, *6*, *13*). In the AFM measurements, the tethered ubiquitins are at relative concentration above the mM range. Therefore, the still-unfolded ubiquitin chains might be expected to associate when the pulling force is reduced, which would produce the kinds of results observed by Fernandez and Li (*1*).

The small number of single-protein folding events observed by Fernandez and Li appear to be barrier-limited. The trajectories [see figure 5 in (*1*)] have a quiescent period followed by a sudden collapse to the native state, the hallmark of a nucleation process. Furthermore, a histogram of the dwell times results in a zero-force extrapolated rate that is within a factor of two of the value observed for barrier-limited folding in solution.

For the single-protein events, the collapse process itself takes 0.1 s. This time scale is orders of magnitude slower than what is anticipated from the solution studies. In solution, post–transition state species do not accumulate. Hence, their lifetimes must be less than a millisecond, the approximate time constant of the entire two-state reaction. Hopefully, further studies will clarify the nature of the slow collapse phase observed in the AFM studies.

T. R. Sosnick
Department of Biochemistry
and Molecular Biology
University of Chicago
Chicago, IL 60637, USA

References

1. J. M. Fernandez, H. Li, *Science* **303**, 1674 (2004).
2. S. Khorasanizadeh, I. D. Peters, T. R. Butt, H. Roder, *Biochemistry* **32**, 7054 (1993).
3. S. Khorasanizadeh, I. D. Peters, H. Roder, *Nature Struct. Biol.* **3**, 193 (1996).
4. M. S. Briggs, H. Roder, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2017 (1992).
5. J. Sabelko, J. Ervin, M. Gruebele, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6031 (1999).
6. B. A. Krantz, L. Mayne, J. Rumbley, S. W. Englander, T. R. Sosnick, *J. Mol. Biol.* **324**, 359 (2002).
7. G. W. Platt, S. A. Simpson, R. Layfield, M. S. Searle, *Biochemistry* **42**, 13762 (2003).
8. C. G. Benitez-Cardoza *et al.*, *Biochemistry* **43**, 5195 (2004).
9. S. T. Gladwin, P. A. Evans, *Fold. Des.* **1**, 407 (1996).
10. T. Sivaraman, C. B. Arrington, A. D. Robertson, *Nature Struct. Biol.* **8**, 331 (2001).
11. B. A. Krantz, R. S. Dothager, T. R. Sosnick, *J. Mol. Biol.* **337**, 463 (2004).
12. J. Jacob, B. Krantz, R. S. Dothager, P. Thiyagarajan, T. R. Sosnick, *J. Mol. Biol.* **338**, 369 (2004).
13. B. A. Krantz, T. R. Sosnick, *Biochemistry* **39**, 11696 (2000).
14. B. A. Krantz, L. B. Moran, A. Kentsis, T. R. Sosnick, *Nature Struct. Biol.* **7**, 62 (2000).
15. H. M. Went, C. G. Benitez-Cardoza, S. E. Jackson, *FEBS Lett.* **567**, 333 (2004).
16. E. A. Lipman, B. Schuler, O. Bakajin, W. A. Eaton, *Science* **301**, 1233 (2003).
17. A. A. Deniz *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5179 (2000).

1 June 2004; accepted 26 August 2004

Response to Comment on "Force-Clamp Spectroscopy Monitors the Folding Trajectory of a Single Protein"

Science moves forward when new techniques uncover unanticipated results, and the field of protein folding is no exception. Indeed, our force-clamp spectroscopy measurements of the folding of ubiquitin chains (*1*) revealed trajectories that departed from the expected two-state folding reactions observed with chemical denaturation techniques (*2*). However, the mechanical and chemical studies of protein folding involve very different endpoints and therefore are not directly comparable. An important difference is that these two experimental approaches result in very different changes in the length of the folding protein. A mechanically stretched and unfolded polyprotein begins its folding trajectory from a well-defined point at which the polypeptide can be extended to the point of losing its secondary structure. For example, at a stretching force of 110 pN, ubiquitin is extended by ~86% of its contour length (*1*, *3*). By contrast, a chemical folding trajectory begins from an unfolded state that is far more compact and less well defined (*4*, *5*). Although the trajectory of a protein that folds after chemical denaturation involves changes in the end-to-end distance of at most a few nanometers (*2*, *6*), force-clamp spectroscopy monitors folding trajectories that can be up to several hundred nanometers in length. Even the more steplike final folding contraction [see figure 5 in (*1*)] of a single ubiquitin involves a reduction in length of more than 15 nm and appears rate-limited.

The asymmetry observed between the stepwise unfolding and the folding trajectories reveals a more complex energy landscape than that monitored by chemical denaturation experiments. This is not surprising, given that extension of the unfolded protein to near its contour length drives the protein much further away from the native state and thereby explores new regions of the folding landscape. From this perspective, the classical view of barrier crossing in protein folding may only apply to small extensions away from the native state (*7*).

This debate also raises the more general question of how relevant the available experimental methods are to in vivo protein folding. In view of the force of gravity and the need of living organisms to perform mechanical work, mechanical stretching is very likely to have played a role in the evolution of proteins. By contrast, the large changes in temperature or chemical denaturants commonly employed in protein-folding studies (*2*) are not found in living cells. Furthermore, chemical or thermal denaturation experiments typically define folding through changes in fluorescence of a tryptophan residue or fluorescence resonance energy transfer (FRET) pairs. Although such measurements provide accurate kinetic information, they do not reveal to what degree the folding proteins have recovered their native form. By contrast, the recovery of mechanical stability monitored by force-clamp spectroscopy (*1*) provides an excellent indication of whether the native state has been reached, given that natively folded proteins exhibit mechanical resistance before unfolding.

Although the mechanical folding trajectories observed by force-clamp spectroscopy still defy explanation, we do not agree with the proposal advanced by Sosnick (*8*) that the folding trajectories of a ubiquitin chain represent the incongruous collapse of aggregating protein modules, driven mostly by their forced intimacy. Simple collapse due to aggregation would not lead to the correct folding of the individual ubiquitins in the chain, which is our main observation. Furthermore, the folding of contiguous protein modules is likely to be a common theme in the function of modular proteins such as titin (*9*), tenascin (*10*), spectrin (*11*), ubiquitin (*3*), and many others. Evolutionary pressure on these proteins must have resulted in mechanisms that effectively avoid the entanglement of folding neighbors (*12*). From this perspective, the mechanical folding trajectories captured by force-clamp spectroscopy reflect much more closely the folding of such modular proteins in vivo, compared with those obtained by means of thermal or chemical manipulations of isolated monomers.

Julio M. Fernandez
Department of Biological Sciences
Columbia University
New York, NY 10027, USA

Hongbin Li
Chemistry Department
University of British Columbia
Vancouver, British Columbia,
Canada V6T 1Z1

Jasna Brujic
Department of Biological Sciences
Columbia University
New York, NY 10027, USA

References

1. J. M. Fernandez, H. Li, *Science* **303**, 1674 (2004).
2. J. Jacob, B. Krantz, R. S. Dothager, P. Thiyagarajan, T. R. Sosnick, *J. Mol. Biol.* **338**, 369 (2004).
3. M. Carrion-Vazquez *et al.*, *Nature Struct. Biol.* **10**, 738 (2003).
4. D. Shortle, M. S. Ackerman, *Science* **293**, 487 (2001).
5. K. W. Plaxco, M. Gross, *Nature Struct. Biol.* **8**, 659 (2001).
6. I. S. Millett, S. Doniach, K. W. Plaxco, *Adv. Protein Chem.* **62**, 241 (2002).
7. M. Carrion-Vazquez *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3694 (1999).
8. T. R. Sosnick, *Science* **306**, 411 (2004); www.sciencemag.org/cgi/content/full/306/411b.
9. M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, H. E. Gaub, *Science* **276**, 1109 (1997).
10. A. F. Oberhauser, P. E. Marszalek, H. P. Erickson, J. M. Fernandez, *Nature* **393**, 181 (1998).
11. M. Rief, J. Pascual, M. Saraste, H. E. Gaub, *J. Mol. Biol.* **286**, 553 (1999).
12. A. F. Oberhauser, P. E. Marszalek, M. Carrion-Vazquez, J. M. Fernandez, *Nature Struct. Biol.* **6**, 1025 (1999).

1 July 2004; accepted 22 September 2004

SCIENCE AND LAW

Science Weighs In on the Scales of Justice

Bettyann Holtzmann Kevles

Science has played an increasingly visible role in the courtrooms of the United States as the benefits and, inevitably, the hazards of medicine and technology affect more and more people. In the 20th century, discoveries about time, space, the structure of matter, and the biology of heredity permeated popular discourse, giving many scientists celebrity status. Not surprisingly, scientists have been called to testify as experts at civil and criminal trials; they have been sought out for their expertise in a wide range of fields, from physics, chemistry, and biology to social sciences such as psychology and sociology. Tal Golan and David L. Faigman have both written original and thoughtful histories of the fitful relationship between science and the law from its roots in the 18th century common law.

In his deft descriptions of a handful of precedent-making cases that elevated the authority of scientific experts in our adversarial legal system, Golan (a historian of science at the University of California, San Diego) is directly concerned with the development of the sciences themselves. Reminding us of the use of experts in English common law, Golan describes a medieval world in which judges sought advice from experts they assumed were impartial. By the 18th century, however, experts were being paid by one side or the other, and the verbal duels we are familiar with began to dominate some civil and criminal cases.

Golan begins *Laws of Men and Laws of Nature* in England with the "origin" case that set the rules governing the use of scientific experts for years to come. In the seaport of Wells on the Norfolk coast, the town commissioners sued two great owners of agricultural land, demanding the removal of embankments that, the commissioners claimed, had destroyed the harbor. The commissioners represented shippers who had grown rich when in the mid-18th century the harbor at Wells had become one of England's busiest ports. By the 1780s, however, it had silted up, forcing ships to dock beyond the marshes. The commission attributed the harbor's decline to the embankments, which had drained the salt marshes, yielding acres of arable land.

Each party hired its own experts. The commissioners selected civil engineers, including a member of the Royal Society. The landowners' principal expert was an engineer and member of the Royal Society, too, but he was also a "natural philosopher"—a scientist—who testified that the harbor had succumbed to the natural forces of wind, sea, and tide. In the first of a series of trials, the judge rejected the scientist's testimony because it was theoretical and not based on hands-on experience with the harbor. But the next year, 1782, a new judge reversed that decision, declaring that natural philosophers who testified about general principles were acceptable experts. This opinion opened the courtroom to scientists as expert witnesses.

After Golan discusses cases that involve chemistry and medicine (poisonings appear frequently in Victorian history), his narrative crosses the Atlantic. He notes that the treatment of scientific evidence by the legal systems in England and the United States diverged when English judges asserted the right, unavailable to Americans, to defuse debates by guiding juries on the relative merits of contending experts.

The next cases that Golan describes, from the late 19th and early 20th centuries, pivot on "machine interpreters," experts hired to explain how newly developed machines provided scientific data. The earliest, an especially grizzly murder trial, involved a microscopist, an expert at examining biological material with recently improved microscopes. The young fraternity of microscopists was interested in establishing credentials. However, while they could identify mammalian blood (samples of which were found on the defendant's clothes), they did not agree about the possibility of distinguishing human blood from that of other mammals. Under these murky circumstances, in 1892 the president of the American Microscopical Society announced his opposition to microscopists giving testimony that could endanger a defendant's life. That did not discourage his colleagues, who enjoyed the pay and continued to testify when asked.

X-rays were used in American trials in 1896 (within a year of their discovery), and protoradiologists argued that courts needed experts to interpret the sometimes-shadowy pictures that resulted from directing the rays through a patient onto a photographic plate. At about this time, Golan explains, Hugo Münsterberg, an experimental psychologist, tried unsuccessfully to convince the courts that a psychological interpretation of mental processes could reveal if a witness was telling the truth. A generation later, William Marston (one of Münsterberg's students) perfected a machine that, connected to a subject, measured physiological changes in response to questioning. In 1922, he offered his "lie detector" to a Washington, D.C., court in the murder trial of James Alphonso Frye. But the judge ruled the machine inadmissible, and the following year the Court of Appeal upheld that ruling on the grounds that the lie detector had not "gained general acceptance in the par-

Laws of Men and Laws of Nature
The History of Scientific Expert Testimony in England and America
by Tal Golan

Harvard University Press, Cambridge, MA, 2004. 335 pp. $49.95, £32.95, €46.10. ISBN 0-674-01286-0.

Laboratory of Justice
The Supreme Court's 200-Year Struggle to Integrate Science and the Law
by David L. Faigman

Times Books, New York, 2004. 432 pp. $27.50, C$41.95. ISBN 0-8050-7274-8.

Justice. This is one of six massive marble statues adorning the façade of the Shelby County Courthouse (1910), Memphis, Tennessee.

CREDIT: KEVIN FLEMING/CORBIS

The reviewer is in the Department of History, Yale University, Post Office Box 208324, New Haven, CT, 06520–8324. E-mail: bettyann.kevles@yale.edu

Published by AAAS

ticular field in which it belongs." This definition of what was scientifically acceptable at a trial, known as the *Frye* rule, dominated the foggy field of scientific testimony in American courts for most of the 20th century. Galon argues that the real reason for the court's rejection of the lie detector was its usurpation of the jurors' right to determine truth. Later, seated juries rejected solid, undisputed scientific evidence (such as blood types in paternity cases and DNA evidence in murders), perhaps in a like resentment of the authority of science.

David Faigman (a professor at Hastings College of Law at the University of California) explores a different part of the legal arena in *Laboratory of Justice*. He, too, reaches back two centuries, but where Golan looks at the trials where legal battles began, Faigman focuses on where selected American cases conclude: the U.S. Supreme Court. Faigman also uses pivotal historical cases, but those he chooses concern decisions that illustrate how Supreme Court justices too often fail to address empirical questions about science.

Faigman describes the intellectual and social milieu that informed the acceptance of suspect scientific ideas of the judges responsible for some of the great miscarriages of justice. To this end, he devotes much of his narrative to descriptions of the families, colleagues, friends, and intellectual disciples of the particular figures he takes to task. Among his targets is Roger Taney, known for his decision in the infamous Dred Scott decision (1857). Faigman finds Taney's thinking rife with a sloppy, lazy disregard of science as well as rich with an unscientific devotion to the intentions of the founding fathers. (For example, they could be said to have favored the institution of slavery, regardless of its morality, because slaves are, indeed, mentioned in the Constitution.) The author is even harder on Oliver Wendell Holmes, who accepted the now-discredited tenets of eugenics without bothering to investigate the facts. Holmes spoke for the majority in the 1927 case of *Buck* v. *Bell*, which condemned a child to sterilization when hearsay called her retarded because her mother and grandmother were slow. He concluded his remarks saying "three generations of imbeciles are enough."

Faigman is particularly critical of decisions justified in the name of science that defend society as a whole at the expense of the rights of the individual. Two prevalent themes in his book are the justices' failures to incorporate scientific knowledge into their reasoning and to recognize that science is a moving target. When complicated issues like the impact of chemicals or new drugs are at issue, he urges the judiciary to engage with the sciences whose products and by-products affect so many people.

Although there is almost no overlap between the cases covered in the books, both authors see a danger in the proliferation of self-serving scientists whose expertise goes to the higher bidder. And both are encouraged by the 1993 Supreme Court decision in *Daubert* v. *Merrell Dow Pharmaceuticals, Inc.* In that case, the lawyers for Daubert, a child with birth defects, attributed his condition to the effects of Bendectin, an antinausea drug taken by his mother. They lost because the court found that there was no acceptable scientific evidence that the drug was at fault. The decision included the instructions to trial court judges to act as gatekeepers in selecting scientific experts. Today, the *Daubert* rules have largely replaced *Frye*. Golan and Feigman each applaud that change as a significant step toward integrating valuable scientific expertise into the judicial system of a society increasingly dependent on the fruits of science.

PSYCHOLOGY

Lessons from Primates

Francine Dolins

"Just how does one listen to a chimpanzee?" This is one of the many nuanced questions Duane Rumbaugh and David Washburn address in *Intelligence of Apes and Other Rational Beings*. The best way to approach primate intelligence, they argue, is to study animals that are afforded opportunities to behave in contexts appropriate to their species, to surpass their immediate training and experience, and to demonstrate creativity and rational behavior. Observations of such animals provide the foundations for the theoretical framework, "rational behaviorism," the authors offer as a new way to understand animal learning and cognition.

Intelligence of Apes and Other Rational Beings
by Duane M. Rumbaugh and David A. Washburn
Yale University Press, New Haven, CT, 2003. 344 pp. $37.50, £29. ISBN 0-300-09983-5. Current Perspectives in Psychology.

The idea that animals display rational behavior has a long history, which includes Darwin's theory of continuity of traits among species. In the course of developing their own theory, the authors (primatologists at Georgia State University) provide an informative survey of this earlier work, from Descartes's view that animals are

The reviewer is in the Department of Psychology, University College Winchester, Winchester, Hampshire SO22 4NR, UK. E-mail: Francine.Dolins@winchester.ac.uk

senseless machines through behaviorist concepts to views currently held in comparative psychology.

Rational behaviorism posits that the intelligent, novel behaviors animals exhibit to achieve specific goals are not learned solely through experience, nor are they shown by all members of a species. Instead, some spontaneously emerge as dynamic responses, more than the sum of their parts, elicited by adaptive challenges. To Rumbaugh and Washburn, these behaviors "resist a conditioning explanation but seem to reflect animals' natural and active inclination to seek predictive relations," which they call "emergents." The process by which behavioral patterns are altered to creatively solve novel problems and the question of how science interprets the origins of such behavior lie at the crux of the book.

According to the authors, Pavlov's respondents (actions elicited by a stimulus) and Skinner's operants (actions that produce a change in the environment) provide bases for emergents. But, in keeping with the authors' perspective of animals as thinking beings, their concept extends well beyond that of the behaviorist's stimulus-response bond. That contingency of associated events is crucial to learning has been agreed upon for nearly a century. However, the idea of emergents, which use relational learning while adhering to basic stimulus-response principles, reflects the flexibility inherent in organisms' responses to ever-changing and challenging environments. In rational behaviorism, instead of learning only specific tasks, animals learn about tasks in relation to their own motivational states and internal goals. As the authors describe, Harlow's learning-set experiments have shown that animals, particularly monkeys, can "learn how to learn" by deriving "hypotheses" or rules about the types of problems they encounter and then applying these to new classes of problems. In addition, captive animals will often work for food rewards but may not even consume those rewards (behavior referred to as "contrafreeloading"). In such situations, where is the reinforcer (the reward)? And what value does it possess in relation to traditional behaviorist views and in the proposed rational behaviorism?

In discussing links between other stimuli and the eliciting properties of the reinforcer, the authors rely on the principles of temporal contiguity and the attention to salient cues in the environment. In sensory preconditioning and other conditioning procedures, the animals do not learn only about the relation among associated temporal events. They also gain information about the types of reinforcers involved in the association as a class or system—knowledge with which they can potentially make relational inferences about novel combinations and novel stimuli. Rumbaugh and Washburn conclude that "the reinforcer essentially is but a salient stimulus that imparts its function in eliciting behavior to other salient stimuli" and that it functions "to inform organisms about contextual resources and how they can be accessed by certain kinds of behavior." In light of this reevaluation of the role of the reinforcer, the authors reconfigure traditional behavioristic principles and thus lay out new challenges for the science of behavior.

Language student. Panzee, a female chimpanzee (*Pan troglodytes*) shown here at age three years, and a bonobo (*P. paniscus*) were raised together in a study that examined their untutored mastery of word-lexigrams and their comprehension of human speech.

Whether one views the flexibility of behavior as "gestaltist" insight or derived from experience—or falling somewhere along the continuum between them—one cannot deny the clever and unexpected responses to challenging situations that some animals have demonstrated. For example, the authors describe an accomplishment of Panzee, a female chimpanzee reared in a study of spontaneous learning. Panzee was shown where a few desired foods were hidden in the woods beyond her outdoor exercise yard. Through gesture and her use of a lexigram board (a keyboard with symbols for representing words and phrases), she recruited a person naïve to the task to go outdoors. Panzee then went out into her yard, from where she used gestures and vocalizations first to direct the person's attention to the locations of the hidden foods and subsequently to retrieve them for her benefit. These behaviors had neither been trained nor previously reinforced, and Panzee's manner of obtaining these hidden and distant foods was totally individual to her.

Although the book's focus is not restricted to primate studies conducted at Georgia State University's Language Research Center, that research forms a central strand in the authors' presentation of rational behaviorism. They recount early work with the chimpanzee Lana, and Rumbaugh's innovative use of the computerized lexigram board to empirically monitor linguistic responses—a productive approach that has been applied to explore animal language, cognition, perception, and sensation in labs around the world. They also describe findings from studies of Kanzi, a bonobo that while very young learned to use the lexigram board without any training or reinforcement. (His comprehension of syntax is well established, whereas his production is somewhat limited.) These include results of spectrographic analyses of Kanzi's vocalizations reported only last year, and the findings provide an enticing view for future research on language.

The book is worth the attention of anyone interested in animal learning and behavior. Its interdisciplinary nature links studies in ethology, neurophysiology, behavior, and cognition through the overarching principle of rational behaviorism: Dynamic, novel behaviors can and do emerge in contexts that extend beyond animals' past experiences and the contingency of reinforced patterns of responses. The concept of emergents helps explain the origins of rational and creative behaviors that diverge from patterns normally exhibited by and expected of animals. Rumbaugh and Washburn's theory increases the sophistication of our understanding of complex behaviors and affords animals a more esteemed position in our world. *Intelligence of Apes and Other Rational Beings* provides a window into the ways that animals are creative. It also demonstrates the warmth that those who study these animals have for being inducted into the mysteries that they hold.

CREDIT: DUANE RUMBAUGH

Published by AAAS

CLIMATE

To Hedge or Not Against an Uncertain Climate Future?

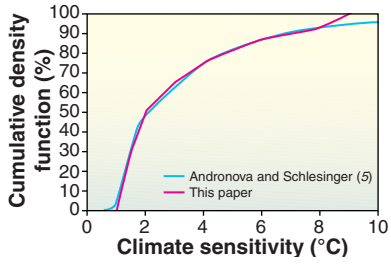
Gary Yohe,[1*] Natasha Andronova,[2] Michael Schlesinger [2]

It has been over a decade since Nordhaus (*1*) published his seminal paper on mitigation policy for climate change. His question was "To slow or not to slow?"; his answer was derived from a traditional cost-benefit approach. He found that a tax levied on fossil fuel in proportion to its carbon content, which would climb over time at roughly the rate of interest, maximized global welfare. Although many more analyses of the same question have since been published, his results are still robust if one assumes a deterministic world in which decision-makers are prescient. However, no decision-maker has perfect foresight, and the uncertainty that clouds our view of the future has led some to argue that near-term mitigation of greenhouse gas emissions would be foolish. Such policy would impose immediate costs, they argue, and have uncertain long-term benefits.

We take a different approach in this Policy Forum by assuming that decision-makers will someday become so concerned about the potential damages associated with climate change that they will take action. Even though it is impossible to determine exactly what sort of mitigation target these future policies might ultimately adopt, a "wait-and-see" approach may no longer be the best near-term policy choice. Should we move soon to intervene in global energy markets as a hedge against the expected cost of meeting a currently unknown policy target?

We follow the modeling approach adopted in the hedging experiments conducted by Manne (*2*) and Yohe (*3*) for the Energy Modeling Forum to explore the policy implications of extreme events. Our analysis is based on a modified version of DICE-99 (Dynamic Integrated Model of Climate and the Economy)—a widely respected model of global economic activity and the damages associated with greenhouse gas–induced temperature change (*4*). We assume that decision-makers evaluate the economic merits of implementing near-term global mitigation policies starting in 2005 that will be in force for 30 years. They know that they will be able to "correct" their policy in 2035, and we assume that decisions will be informed by perfect information about both the climate sensitivity and the policy target. Their goal will be to maximize the expected discounted value of gross world product (GWP, the global equivalent of gross domestic product) across the range of options that will be available at that time (see online material for details and definitions).

The uncertainty in our understanding of the climate system against which these policies will be framed is portrayed in the figure (below). It shows a continuous cumulative distribution function (CDF) of climate sensitivity estimated by Andronova and Schlesinger (*5*) (where climate sensitivity is the temperature increase that results from a doubling of atmospheric concentration of greenhouse gases relative to preindustrial levels). It also shows a version of the same CDF that allowed us, for reasons of practicality, to work with a limited number of sensitivities that were nonetheless representative of the continuous CDF. Each sensitivity is associated with a probability, so that it conformed with the continuous version. Both representations show that climate sensitivities as high as 9°C are possible.



Cumulative distributions of climate sensitivity (*5*).

Several structural and calibration modifications of the DICE-99 model were required to accommodate the wide range displayed in the figure. Because responding to high sensitivities could be expected to put enormous pressure on the consumption of fossil fuel, for example, we limited the rate at which the global economy could "decarbonize" itself (i.e., reduce the ratio of carbon emissions to global economic output) to 1.5% per year.

Calibrating the DICE-99 model to alternative climate sensitivities that span the range displayed in the figure was more involved, because the DICE model includes a parameter that reflects the inverse thermal capacity of the atmospheric layer and the upper oceans. Larger climate sensitivities were associated with smaller inverse capacity values, so that the model could match observed temperature data when run in the historical past. The parameter was defined from optimization of the global temperature departures calculated by DICE and calibrated against the observed departures from Jones and Moberg (*6*) for the prescribed range of the climate sensitivities from 1.5° to 9°C (*7*).

Modest near-term mitigation would maximize discounted GWP, even if no mitigation was done after 2035 (see the supporting online text). Achieving optimality or even meeting specific concentration targets would not, however, necessarily hold temperatures below the 2° to 3° range identified by Smith and Hitz (*8*) and the Intergovernmental Panel on Climate Change (IPCC) (9), as a threshold above which damages caused by gradual climate change would climb dramatically, and by Schneider (*10*) and the IPCC (*9*), as a threshold above which abrupt changes become much more likely. We therefore focused our attention on mitigation pathways designed to limit temperature increases to four targeted levels (recorded in the first row of the table, next page) that straddle this critical threshold.

We assumed that global policy-makers would choose among these options in 2035, when the true climate sensitivity would be revealed; but each target was assumed to be equally likely for the purposes of setting near-term policy in 2005. Maximum discounted GWP was computed using the modified DICE-99 framework for initial 2005 taxes ranging from $0 to $50 per ton of carbon. Some combinations involved doing too little in the near term, so GWP fell as downstream mitigation "ramped-up" to achieve the prescribed temperature limit. Other combinations involved doing too much in the near term, so GWP again fell even though mitigation could be "turned down" after 2035. An ini-

[1]The Department of Economics, Wesleyan University, Middletown, CT 06459, USA. [2]The Climate Research Group, Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

*To whom correspondence should be addressed. E-mail: gyohe@wesleyan.edu

tial tax of roughly $10 per ton of carbon (about 5¢ for a gallon of gasoline that would grow at the rate of interest over time) balances these two sources of loss to maximize expected GWP.

Comparisons drawn from the DICE model across the requisite adjustments for the $10 initial tax and for a wait-and-see policy in a "robustness" chart are displayed in the table (*11*). The second column shows that a 2° target could not be achieved, even if mitigation policy began in 2005, for climate sensitivities above 3°; they are "impossible now" in the parlance of the table. Second, 2° and 2.5° targets could not be achieved if an initial $10 tax policy were imposed in 2005 for climate sensitivities above 4° and 6°, respectively ("impossible later" in the table). Doing nothing through 2035 would put 3° beyond the range of possibility if the climate sensitivity were 7° or higher.

DISCOUNTED ADJUSTMENT COSTS ($) GIVEN AN INITIAL TAX OF $10

Climate sensitivity (degrees)	Temperature target (degrees)			
	2	2.5	3	3.5
1.5	$0	$0	$0	$0
2	$2	$1	$0	$0
3	$4	$3	$1	$0
4	IL	$6	$2	$0
5	IL	$12	$3	$0
6	IN	IL	$4	$1
7	IN	IL	$6	$3
8	IN	IL	$9	$5
9	IN	IL	$12	$9

DISCOUNTED ADJUSTMENT COSTS ($) GIVEN NO MITIGATION THROUGH 2035

Climate sensitivity (degrees)	Temperature target (degrees)			
	2	2.5	3	3.5
1.5	$32	$11	$3	$0
2	$38	$22	$16	$4
3	$180	$29	$18	$22
4	IL	$60	$24	$24
5	IL	$142	$25	$25
6	IN	IL	$27	$28
7	IN	IL	IL	$34
8	IN	IL	IL	$35
9	IN	IL	IL	$38

Implementing near-term mitigation policy versus no mitigation of carbon. Comparing the robustness of implementing near-term mitigation policy through 2035 beginning with an initial tax of $10 per ton of carbon (rising to nearly $33 per ton in 2035) with the robustness of imposing no mitigation policy through 2035. Values report losses in discounted GWP (in billions of dollars) when the indicated near-term policy is compared with the minimum-cost deterministic path. Annual losses (and gains) are discounted back to 2005 (see the supporting material on *Science* online). "IN" means "impossible now"; i.e., that the indicated temperature target cannot be reached by any mitigation policy initiated in 2005. "IL" means "impossible later"; i.e., that the indicated targets could not be achieved by any adjustments in 2035 to the specified near-term interventions in 2005.

An initial $10 tax policy is remarkably robust across the remaining possibilities, as shown in the table. Discounted adjustment costs are smaller than $10 billion except for high climate sensitivities near the border of the impossibility frontier. A wait-and-see approach leaves the global economy open for far more serious adjustment costs. Except for higher targets with low sensitivities, doing nothing through 2035 imposes costs in excess of $20 billion in more than half of the possible cases and significantly larger than $50 billion for low temperature targets even with lower climate sensitivities (*12*). These costs are comparable, for example, to the estimated cost of rebuilding Iraq.

We need to be clear that the initial tax would climb over time, as in the original Nordhaus paper (*1*), at the rate of interest. Although some energy sectors around the world might not respond significantly to the initial $10 intervention, the model also captures more vigorous responses in subsequent years—the results of additional incentives created by persistent and growing carbon taxes designed to punish those who ignore conservation and substitution opportunities.

It should not be a surprise that hedging is a preferred strategy in a world where a temperature target may be selected sometime in the future. People buy insurance against extreme events when the risks affect private property, and societies require insurance when potential losses are distributed across a population. It is, however, surprising that climate insurance over the near term can be so inexpensive and that an economically efficient near-term hedging policy can be so robust across a wide range of futures in comparison with doing nothing. The point is that paralysis in near-term action can make temperature targets as low as 3° impossible to achieve if the climate sensitivity turns out to be higher than 6°. Moreover, the cost of adjustment measured in terms of discounted GWP can be many times higher for lower climate sensitivities if nothing were done for 30 years. In short, taking an insurance approach to the near-term mitigation question strongly supports starting modest but persistent intervention on a global scale as soon as possible.

The specific cost estimates are, of course, highly dependent on the global modeling context of the DICE-99 model, the analytical decision to include only uncertainty about climate sensitivity in the analysis, and the identified boundaries of the "impossibility frontier"; i.e., the temperature limits that could not be achieved now and others that could not be achieved if mitigation were delayed for 30 years. In addition, it is highly unlikely that many (if any) of the fundamental uncertainties associated with the climate problem will be resolved over the next 30 years. As a result, we should expect that "midcourse" corrections will involve repeated hedging exercises and thus, relative to the modeling framework presented here, more uncertainty. The qualitative conclusion supporting modest near-term mitigation is, nonetheless, extremely robust, because it is uncertainty that produces its value. Adding other sources of uncertainty would simply add to that value by widening the range of futures over which we must hedge. Uncertainty is the reason for acting in the near term, and that uncertainty cannot be used as a justification for doing nothing.

References and Notes

1. W. D. Nordhaus, *Econ. J.* **101**, 920 (1991).
2. A. S. Manne, "A summary of poll results: EMF 14 Subgroup on Uncertainty" (Stanford Univ., Stanford, CA, 1995).
3. G. Yohe, *Glob. Environ. Change* **6**, 87 (1996).
4. W. D. Nordhaus, J. Boyer, *Warming the World: Economic Models of Global Warming* (MIT Press, Cambridge, MA, 2001).
5. N. G. Andronova, M. E. Schlesinger, *J. Geophys. Res.* **106** (D19), 22605 (2001).
6. P. D. Jones, A. Moberg, *J. Climate* **16**, 206 (2003).
7. Table S1 of the supporting material provides the precise calibration of the CDF for climate sensitivity.
8. J. Smith, S. Hitz, "Estimating the global impact of climate change" [ENV/EPOC/GSP(2003)12, Organization for Economic Co-operation and Development (OECD), Paris, 2003].
9. Intergovernmental Panel on Climate Change (IPCC), *Climate Change 2001: Impacts, Adaptation and Vulnerability* (Cambridge Univ. Press, Cambridge, 2001), chapter 19.
10. S. Schneider, "Abrupt non-linear climate change, irreversibility and surprise" (ENV/EPOC/GSP(2003)13, OECD, Paris, 2003).
11. R. J. Lempert, M. E. Schlesinger, *Clim. Change* **45**, 387 (2000).
12. These costs represent only the added expense of having been wrong in setting mitigation policy relative to the ultimate resolution of the temperature target and of uncertainty about climate sensitivity. The supporting online material records net benefits (using perfect knowledge in 2005 as a baseline) to show that the discounted costs of achieving specific temperature targets can be much larger than these adjustment costs.

13. G.Y. was supported by NSF through its funding of the Center for Integrated Study of the Human Dimensions of Global Change at Carnegie Mellon University under Cooperative Agreement SBR 95-21914. N.A. and M.S. were supported by NSF under Award No. ATM-008420. Any opinions, findings, and conclusions expressed in this publication are those of the authors and do not necessarily reflect the views of NSF.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5695/416/DC1

MATERIALS SCIENCE

Some Assembly Required

Sharon C. Glotzer

Self-assembly—the spontaneous organization of matter into ordered arrangements—is a governing principle by which materials form (*1*). The patterns arising from self-assembly are ubiquitous in nature, from the opalescent inner surface of the abalone shell to the internal compartments of a living cell. Much of materials science and soft condensed-matter physics in the past century involved the study of self-assembly of fundamental building blocks (typically atoms, molecules, macromolecules, and colloidal particles) into bulk thermodynamic phases. Today, the extent to which these building blocks can be engineered has undergone a quantum leap. We are on the verge of a materials revolution in which entirely new classes of "supermolecules" and particles will be designed and fabricated with desired features, including programmable instructions for assembly. These new building blocks will be the "atoms" and "molecules" of tomorrow's materials, self-assembling into novel structures made possible solely by their unique design.

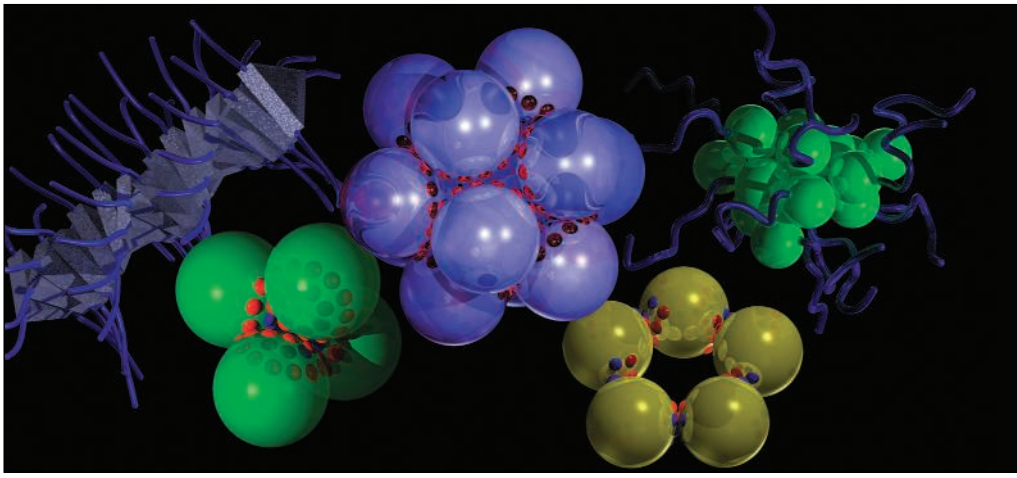
What happens when traditional atoms and molecules are replaced with these new building blocks? What types of ordered structures are possible, and what unique properties do they have?

Colloidal polyhedra (*2*), nanocrystals in the form of tetrapods (*3*) and triangles (*4*), and tiny cubes of molecular silica (*5*) are just a few examples of new building blocks being made today. In most cases, these building blocks may not naturally assemble into any desired structures. One emerging approach to confer upon nanoparticles and colloids predetermined "instructions" for assembly is to decorate the surface of the particles with "sticky patches," made, for example, of synthetic organic or biological molecules. This strategy takes its inspiration in part from biology, where the precision of self-assembled structures such as viruses and organelles originates in the selectivity of the interactions between their constituents. According to computer simulations, synthetic "patchy particles" should self-assemble under the right conditions into structures atypical of traditional materials (*6*) (see the figure).

On macroscopic scales, millimeter-sized plastic wedges patterned with patches of solder and hydrophobic lubricant self-assemble under surface tension when dispersed in water to form tiny electronic devices whose structure resembles that of the tobacco mosaic virus (*7*). Making patchy particles with precise patterns of interactions on nanometer scales is much more challenging, but exciting developments are being reported. For example, Stellacci and co-workers (*8*) recently synthesized gold and silver particles 4 nm in diameter, using organic molecules to control the size of the nanoparticles. Although the use of organic stabilizing layers is commonplace in nanoparticle synthesis, these researchers used a mixture of ligands that, on flat surfaces, would tend to phase separate into bulk phases or random domains. Instead, the ligands self-organized on the nanoparticle surface into repeating patterns of stripes and dots with spacings as small as 0.5 nm, imparting a controllable, precise, and unprecedentedly small pattern of attractive and repulsive patches to the surfaces of the particles. Striped spheres and spheres with polar patches were obtained, providing a striking demonstration of the role of curvature in pattern formation (*9*). This method suggests an exciting strategy for controlling the symmetry of nanoparticle assemblies through anisotropic interactions achieved by patterning. In another example, Mokari *et al.* recently patterned semiconductor tetrapods and nanorods with gold patches on the tips (*10*), potentially providing a new way to assemble components for nanocomputing devices.

Genetic engineering of biomolecules like DNA and proteins opens up further possibilities for conferring recognition (*11*) and chemical specificity to particles, creating building blocks that are potentially capable of assembling into hierarchically arranged structures. In a recent twist, a new patchy particle was synthesized by precisely positioning gold particles onto specific sites on the surface of the cowpea mosaic virus, creating a new type of building block with the potential for self-assembly (*12*).

Patchy particles are but one example of "shape amphiphiles"—building blocks of potentially complex shapes with competing interactions that expand the range of self-assembled structures beyond those exhibited by traditional amphiphiles such as surfactants and block copolymers. By attaching polymeric "tethers" to nanoparticles, another new class of shape amphiphile may be fabricated (*13*). These building blocks can



Predicted self-assembled structures for model building blocks. When selective interactions are introduced to particle surfaces through patterning of ligands or polymeric tethers, competing interactions can cause the particles to self-organize into complex structures (*6*, *13*). (**Left**) Twisted wire of tethered triangular nanoparticles; (**middle**) tetrahedron, icosahedron, and ring self-assembled from spherical patchy particles; (**right**) micelle of tethered nanospheres. To fabricate rings from patchy particles, selective sticky patches are placed anisotropically on the equatorial plane at a relative angle of <180°. The diameter of the rings is controlled by the angle between the patches. Tetrahedra and icosahedra form from particles with selective, ringlike patches shifted off the equatorial plane.

CREDIT: C. R. IACOVELLA/UNIVERSITY OF MICHIGAN

The author is in the Department of Chemical Engineering and the Department of Materials Science and Engineering, University of Michigan, 2300 Hayward Street, Ann Arbor, MI 48109–2136, USA. E-mail: sglotzer@umich.edu

Published by AAAS

form structures that combine the features of self-assembling surfactant or block copolymer systems with the intricate ordered phases of liquid crystals (see the figure). Patterning techniques such as that described above may provide a means to position tethers at specific locations on the particle surface. If this can be achieved, simulations predict that the combination of forces, particle shapes, and building-block topology will provide a means for assembling the particles into wires, sheets, tubes, and other structures. Examples of tethered building blocks already synthesized include poly(ethylene glycol)–tethered CdTe quantum dots (*14*), poly(ethylene oxide)–tethered fullerenes (*15*), and PEG-tethered silica cubes (*16*). Many more are sure to follow.

In contrast to traditional materials, where materials are selected, rather than designed, for specific applications, the next generation of materials will benefit from the a priori design of novel building blocks, programmed for assembly and synthesized with particular needs in mind. With the rapid pace of developments in this field, humankind's newest atoms and molecules are just around the corner.

References and Notes

1. G. M. Whitesides, M Boncheva, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4769 (2002).
2. V. N. Manoharan *et al.*, *Science* **301**, 483 (2003).
3. D. J. Milliron *et al.*, *Nature* **430**, 190 (2004).
4. N. Malikova *et al.*, *Langmuir* **18**, 3694 (2002).
5. R. M. Laine *et al.*, *J. Appl. Organomet. Chem.* **12**, 715 (1998).
6. Z. L. Zhang, S. C. Glotzer, *Nano Lett.* **4**, 1407 (2004).
7. D. H. Gracias *et al.*, *Appl. Phys. Lett.* **80**, 2802 (2002).
8. A. M. Jackson *et al.*, *Nature Mater.* **3**, 330 (2004).
9. D. R. Nelson, *Nano Lett.* **2**, 1125 (2002).
10. T. Mokari *et al.*, *Science* **304**, 1787 (2004).
11. S. Wang *et al.*, *Nano Lett.* **2**, 817 (2002).
12. A. Szuchmacher Blum *et al.*, *Nano Lett.* **4**, 867 (2004).
13. Z. L. Zhang *et al.*, *Nano Lett.* **3**, 1341 (2003).
14. S. Westenhoff, N. A. Kotov, *J. Am. Chem. Soc.* **124**, 2448 (2002).
15. T. Song *et al.*, *Polymer* **44**, 2529 (2003).
16. G. Cardoen, E. B. Coughlin, *Macromolecules* **37**, 5123 (2004).

17. Supported by the NSF (grants CTS-0210551 and DMR-0103399) and U.S. Department of Energy (grants DE-FG02-02ER46000 and DE-FG02-03-ER46094).

ECOLOGY AND CONSERVATION

Space—The Final Frontier for Economists and Elephants

Erwin Bulte, Richard Damania, Lindsey Gillson, Keith Lindsay

At the Convention on International Trade in Endangered Species this month (*1*), Namibia is asking for an annual quota for the sale of ivory that is "accumulated from natural and management-related mortalities." The discussion is likely to be steeped in controversy, not least because of the complexity of the economic and ecological arguments involved. Managing elephant populations and evaluating the sustainability of the ivory trade require not only detailed economic analyses, but also recognition of the ecological complexities that influence decisions about elephant management.

Enhanced online at www.sciencemag.org/cgi/content/full/306/5695/420

Understanding the economics of natural resources is crucial in such policy deliberations. So-called bioeconomic modeling—which describes interactions between commodity markets and biological populations such as elephant populations—has provided useful insights into two principal aspects of the ivory trade. First, bioeconomic modeling has shown that poaching and legal harvesting of ivory are not independent, although the nature of this interrelationship is still disputed. Some economists argue that banning a legal ivory trade might give an impetus to the black market and boost poaching (*2*). Others suggest that legal harvesting and trade may facilitate the "laundering" of illegal products—a potentially important but untested hypothesis (*3*). Second, economists have debated the effects that revenues from the ivory trade might have on conservation. On the one hand, it can be argued that ivory sales might provide incentives for governments to carefully manage the resource. For example, governments may be encouraged to invest in the monitoring of elephant populations, to enforce laws against illegal hunting and poaching, and to set aside land as elephant habitat [the species "earns its way" (*4*)]. In the absence of such revenues, with growing elephant and human populations competing for land, it has been pointed out that wildlife may be exploited unsustainably, and that habitat will be converted to other more lucrative purposes by local people or investors. Conversely, recent developments in political economics emphasize that high commodity prices for ivory may be bad for conservation. High prices may unleash forms of "rent seizing" and patronage politics whereby vested interests seek to dismantle the protective institutions that limit their ability to grab the resource (*5*). Notwithstanding these contributions and the conflicting signals they send, economic models of elephant management and the ivory trade have failed to capture several essential elements.

From an economic standpoint, the simplified treatment of the roles played by national and international institutions and the fact that most ivory trade models ignore feedback from other land and labor sectors of national economies suggest that these models are incomplete. These are important omissions. A recent study reveals an association among poor governance, corruption, and declining elephant populations (*6*). Brander and Taylor (*7*) emphasize that incompletely enforced property rights (as is evidently the case for many elephant populations) and a relaxation of ivory trade controls may not only be detrimental for conservation, but also may reduce human welfare in countries where elephants roam (the "range states"). In particular, given that resuming legal trade may have uncertain effects on ivory market prices, it is unclear how incentives to poach will be affected in range states that export ivory and possibly in range states that do not trade in either African or Asian elephant products (an external effect).

Large-bodied species like elephants have slow population growth rates and are particularly at risk from overexploitation. As benefits from tourism are positively affected by the size of elephant populations and negatively affected by poaching mortality and the enforcement costs needed to protect elephants, the net benefits of resuming the ivory trade are inherently uncertain. Regulated

CREDIT: ORLA SMITH/*SCIENCE*

E. Bulte is in the Department of Economics, Tilburg University, Post Office Box 90153, 5000 LE Tilburg, Netherlands. R. Damania is in the School of Economics, University of Adelaide, Adelaide 5005, Australia. L. Gillson is in the Environmental Change Institute, Biodiversity Research Group, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. K. Lindsay is with the Amboseli Elephant Research Project, Amboseli Trust for Elephants, Post Office Box 15135, Langata 00519, Nairobi, Kenya.

trade may be preferable to free trade, and the optimal level of regulation "stringency" (a trade ban, sale of ivory stockpiles, or regulated sale of ivory harvested from wild populations) depends on factors that are as yet poorly researched or understood. Furthermore, net revenues from ivory sales often are only part of the income generated from elephants. Such income may also include photo tourism and sports hunting, although this depends on local circumstances (*8*). Bioeconomic models of the ivory trade should be augmented to capture these complex issues.

Current economic models are ecologically simplistic because they are underpinned by the convenient though often false idea of "equilibrium." Until recently, deterministic single-species models, which do not consider the interactions between a particular species and a variable multispecies environment, have dominated the field. Most economic analyses of the ivory trade are based on a simple logistic model (*4*, *9*), which assumes that the population growth rate for elephants will decline until it reaches zero when elephant numbers reach the "carrying capacity" of the environment. At the "carrying capacity," the size of a population will, in theory, remain constant, because birthrate and death rate are equal and environmental resources (such as forage) are consumed at the same rate as they are produced. Such models do not recognize that elephant habitat may not be at equilibrium with climate, and that rainfall and forage abundance vary on time scales that range from years to decades (*10*). This variability in turn affects reproductive and mortality rates, the age structure of elephant populations, and hence the supply of ivory over time. Moreover, the "quality" of habitat, or the ability of land to support wildlife, is also affected by other factors—elephant density, fire, and economic activities at the margins of protected areas. Rather than sitting at a single idealized "carrying capacity," ecosystems may occupy one of several "multiple stable states" at any given point in time or may be in transition from one state to another (*11*).

Although some economists have attempted to focus on multispecies models (*12*) and models that incorporate temporal variability (*13*) and spatial scale (*14*), there remains a gap between stylized economic models and recent ecological thinking. There is a strong need to bridge the gap between ecological theory and the economics of natural resources by incorporating variability, complexity, scale, and uncertainty into current economic models.

Space is now widely considered to be the new frontier in environmental and resource economics. Economists are already using theories from the dynamics of fragmented populations ("metapopulations") to analyze the spatial pattern of dispersal and harvesting in marine environments (*14*). Yet when considering terrestrial ecosystems, economists have yet to realize that these are a mosaic of potential interacting sites whose populations cannot be described by simple logistic economic models. Important conservation concerns, currently undervalued by economists, ensure ecological integrity and the maintenance of ecosystems. Dealing with these concerns requires attention to the interactions between the ecological variability of habitat and the economic (opportunity) costs of protecting interconnected habitat patches. Managing and conserving ecosystems in flux differs from older approaches aimed at maintaining stability (*15*). In an ecosystem in flux, population sizes and the distribution of animals and plants vary over time and space. This variability has implications for conservation strategies and management of natural resources, because a constant supply of goods like ivory cannot be guaranteed. Furthermore, the land surrounding the periphery of a protected area should be viewed as equal in importance to the core, because potentially it can accommodate changing distributions of plants and animals. Such a view would help to maintain viable metapopulations across a landscape, preventing animal reserves from becoming isolated and possibly overcrowded or impoverished.

Incorporating the spatial dimension of ecology into economic models permits a more accurate evaluation of the ecological impact and economic costs of alternative policies. Currently, however, there is a mismatch between state-of-the-art economic and ecological theory on the one hand, and the contributions of economists to the debate on the ivory trade on the other. The way forward in the immediate future may be for ecologists to identify the scales at which equilibrium models provide an approximation of reality, and for economists to build this spatial scale into their models. A future goal for economists when analyzing the exploitation of flora and fauna will be to develop models that capture the nonequilibrium nature of ecological systems.

References

1. Convention on International Trade in Endangered Species, 13th Conference of Parties, Bangkok, Thailand, 2 to 14 October 2004.
2. E. B. Barbier *et al.*, *Elephants, Economics and Ivory* (Earthscan, London, 1990).
3. C. Fischer, *J. Environ. Econ. Manage.* **48**, 926 (2004).
4. T. M. Swanson, *Oxford Econ. Pap.* **46**, 800 (1994).
5. M. Ross, *Timber Booms and Institutional Breakdown in Southeast Asia* (Cambridge Univ. Press, Cambridge, 2001).
6. R. J. Smith *et al.*, *Nature* **426**, 67 (2003).
7. J. Brander, M. S. Taylor, *Can. J. Econ.* **50**, 526 (1997).
8. J. Barnes, in *The Future of Botswana's Elephants*, P. Hancock, Ed. (Kalahari Conservation Society, Gaborone, 1990), pp. 60–66.
9. E. H. Bulte, G. C. van Kooten, *Am. J. Agric. Econ.* **81**, 453 (1999).
10. C. B. Barrett, P. Arcese, *Land Econ.* **74**, 449 (1998).
11. H. Dublin *et al.*, *J. Anim. Ecol.* **59**, 1147 (1990).
12. D. Finnoff, J. Tschirhart, *J. Environ. Econ. Manage.* **45**, 589 (2003).
13. J. D. Saphores, *J. Econ. Dyn. Control* **28**, 509 (2003).
14. J. N. Sanchirico, J. E. Wilen, *J. Environ. Econ. Manage.* **37**, 129 (1999).
15. K. H. Rogers, in *The Kruger Experience: Ecology and Management of Savanna Heterogeneity*, J. T. du Toit, K. H. Rogers, H. C. Biggs, Eds. (Island, Washington, DC, 2003), pp. 41–58.

BEHAVIOR

A Marketplace in the Brain?

George Ainslie and John Monterosso

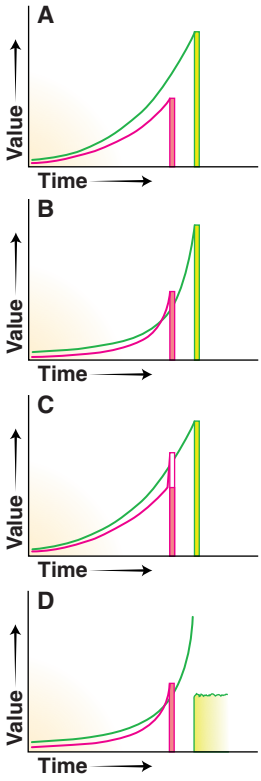
There are a number of studies that investigate violations of rationality in human decision making. One important violation that is repeatedly observed is a tendency to discount expected outcomes proportionate to their delay. This results in a systematic inconsistency of preference over time. On page 503 of this issue, McClure *et al.* (*1*) present an elegant functional magnetic resonance imaging (fMRI) study that measures changes in neural activity as human volunteers are presented with the possibility of delayed rewards. This work is an important step toward direct observation of the decision-making process, although its findings are open to different interpretations.

G. Ainslie is at the Coatesville Veterans Affairs Medical Center, Coatesville, PA 19320, USA. E-mail: george.ainslie@med.va.gov J. Monterosso is in the Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, CA 90024, USA.

The dominant theory in the behavioral sciences has been that normal people discount the option of a delayed reward according to an exponential curve, that is, by a constant percentage per unit time. This exponential curve is similar to that used by financial markets: current value = nondelayed value × (1 – discount rate)$^{\text{delay}}$. Exponential discounting implies a stability of preference over time. Individuals who exhibit exponential discounting behavior when faced with a choice between a smaller/sooner reward or a larger/later reward do not change their preference as the smaller reward becomes imminent. Rather, such individuals continually choose options that maximize their long-range prospects with allowances for the reduced value of the delayed goods. But despite its simplifying appeal, exponential discounting and the implied consistency of preference is not a tenable description of the way that either humans or nonhumans tend to evaluate the future. On the contrary,

there are conditions under which most subjects reverse their initial preference for larger/later rewards when smaller/sooner alternatives are near at hand. Furthermore, when given the opportunity, both human and nonhuman subjects choose to lock in their larger/later preference while this is still the most appealing (dominant) option (*2*–*4*). Thus, behavioral discounting data followed over time look systematically nonexponential. Consistently, such data are better explained by a function in which value varies proportionally with delay, although opinions differ as to the precise form of this function. Statistical methods of fitting the curve to data strongly favor a hyperbolic function as follows: Value = nondelayed value/[1 + (discount rate × delay)] (*5*). The steepness of the proportional curve of value as a function of the time delay varies over a narrow range in nonhuman species but widely in humans (*6*, *7*). Yet the same proportionality of discount rate with delay has been observed over a time period that ranges from seconds to decades, suggesting that the form of this curve may reflect a basic mechanism for perceiving value (*8*). David Laibson, an economist studying this problem and a co-author of the new study, has pointed out that a two-factor discount function, which he calls hyperboloid, can explain preference reversal and choice of prior commitments while retaining an exponential shape for most of the length of the curve (*4*). Laibson's formula is a conventional exponential discount function δ[δ = (1 – discount rate); δ factor value = δ^{delay}] multiplied by a delay penalty factor β [$0 < \beta < 1$ for all delays > 0; thus, value = value if immediate × β × δ^{delay})].

Either form of nonexponential discounting, with its implied conflict between the current preference and the predictable preference of future selves, opens the study of inconsistent choice to bargaining theory. Thus, the whole self can be seen as a series of choice-makers, each influenced differently by the range of future options and each in partial conflict with the others. In the choice situation depicted by panels B, C, and D of the figure, an individual at an early time prefers the larger/later reward, but will need to influence or forestall the future self who prefers the smaller/sooner reward as this reward becomes imminent. The commonsense solution—to use willpower—has always lacked a scientific explanation, and exponential discounting theory does not recognize a reason for why willpower should even be needed. However, with nonexponential discounting, bargaining theory predicts that a person's mere identification of these smaller/sooner versus larger/later choices will bundle the expected rewards into a greater aggregate incentive to pick the larger/later option. In simple terms, individual pieces of chocolate may be irresistible unless the credibility of one's diet is staked against each piece. This sounds like commonsense, but exponential discount curves predict no such bundling effect. Because delay-proportional discount curves decline more slowly as the delay in reward becomes longer and longer, adding them together for a series of rewards should increase differentially the value of larger/later rewards. In fact, this has been observed in experiments with both human (*9*) and nonhuman subjects (*10*). Delay-proportional discounting also predicts temporary preferences at shorter time scales in which the period of dominance of smaller/sooner rewards is too brief to support changes in motor behavior, but is long enough to support shifts in attention. It has been argued (*11*) that these mechanisms provide an instrumental (reward-based) account of phenomena traditionally considered noninstrumental, such as the sudden development of craving induced by an evocative stimulus.



Choosing between immediate and delayed rewards. (**A**) Exponential curves depicting the value of two alternative expected rewards given at discrete time points (the usual set up of behavioral experiments). Smaller/sooner rewards are depicted in pink, larger/later rewards in yellow. (**B**) Hyperbolic curves of the value of two alternative expected rewards showing a temporary preference for the smaller/sooner reward as it draws close. (**C**) β-δ curves of value: The immediacy factor adds a spike to the smaller/sooner curve in (A) to produce a temporary preference. (**D**) Hyperbolic curves of value, with the larger/later curve summed from an extended reward, which is likely to be the case in ordinary life.

In the new work, McClure *et al.* conduct an fMRI study of neural activity in response to a series of rewards of different values and offered with different time delays. Princeton undergraduates were given a series of choices between smaller/sooner and larger/later rewards: The rewards were gift certificates for Amazon.com, ranging from $5 to $40 in value. The smaller/sooner option could be received the same day ("today") or with a 2- or 4-week delay; the larger/later option could be received either 2 or 4 weeks after the smaller/sooner option. The order of presentation was randomized. Subjects did actually receive one of their certificates for the delay chosen, but did not find out which one until after the end of the test. When these subjects were given a choice in which the smaller/sooner option would be delivered the same day, greater activity was observed in corticolimbic regions (ventral striatum and medial orbitofrontal cortex) compared with baseline or with choices not involving a "today" option. The limbic area is known to be active as rewards are anticipated or delivered (*12*–*14*) and in response to emotion-evoking events (*15*). In contrast, relative to baseline, all choices recruited observable activity within the lateral prefrontal cortex and within the parietal cortex in areas associated with future planning (*16*, *17*). When subjects chose a larger/later alternative, there was also greater activity in the lateral prefrontal and parietal areas than when they chose the alternative smaller/sooner reward.

The investigators interpret these findings as evidence for Laibson's β-δ dual discount function, with corticolimbic activity furnishing the β component and the lateral prefrontal activity the δ component. They propose that humans share with nonhumans a lower automatic process governed by the limbic system that motivates impatient emotional choices. This process competes with a uniquely human capacity for general reasoning and future planning that is governed by the lateral prefrontal cortex. The authors argue that sudden elicitation of limbic activity by near-term opportunities or other factors creates the spike in the discount curve that makes it seem hyperboloid and is responsible for temporary reversals of preference. They further suggest that limbic-based cue-conditioned appetites—which Loewenstein has described as "visceral factors" (*18*)—can impose reward contingencies similar to those of reward immediacy. A conditioned stimulus that elicits limbic activation, such as the sight of a tempting dessert or an addict's drug paraphernalia, would presumably remove the β penalty factor at that moment. However, the McClure model does not make it clear whether reward would be discounted for any remaining delay by the rational δ factor or how nonhumans (which are said to lack the δ factor) would discount delayed rewards, much less protect them by observed commitments when these commitments are offered (*2*, *3*).

The differential activation of limbic structures could cause a spike in an otherwise exponential discount curve, but it is the hyperbolic pattern that has been observed in extensive research (*5*). Something about suggesting the possibility of having the

Amazon.com certificate "today" was clearly evocative to the students, but it is not possible to say whether the differential effect of "today" was caused by a surge in anticipated enjoyment of the Amazon books or by the suggestion of winning a more or less immediate prize per se. Nor would it be possible without a series of shorter time delays, ideally of visceral rewards (not money or certificates for later exchange as in the McClure *et al*. study), to tell whether the "β" (limbic) activity is best described as an either/or phenomenon (immediate-yes versus delayed-no) as the authors suggest, or as part of a smooth discount curve that cannot be detected by current fMRI methods for delays of 2 weeks. Although the increased activity of "δ brain areas" (the lateral prefrontal cortex and associated structures) in response to larger/later selections is an important finding, to accord these areas status as a separate decision-making mechanism would add a complicating factor that would have to be reintegrated with motivation. As the behavioral neurobiologists Montague and Berns point out, all organisms need "an internal currency that can be used as a common scale to value diverse behavioral acts and sensory stimuli" (*19*). It may be that the δ brain areas reported by McClure *et al*., in effect, only broker limbic-based rewards. Such a limitation of δ areas was anticipated by the British empiricist David Hume who wrote: "[Reason alone] is incapable of preventing volition. . . . Reason is and ought only to be the slave of the passions" (*20*).

As for cue conditioning, it is at first glance a simpler explanation for sudden craving than is a change in the prospects of the success of a reward-governed appetite. The notion of the β factor certainly has intuitive appeal. But conditioning is now thought to associate stimuli only with other stimuli, not responses (*21*), and thus cannot be the means of transferring reflexive responses from one stimulus to another as was originally thought (*22*). Cue conditioning and the dual β-δ motivational model are largely compatible with the predictions of hyperbolic discounting theory, but they represent additional mechanisms that are probably not needed to fit the data, including the data that McClure *et al*. report. The study discussed here is the first step in an important direction, but is not yet enough to specify the mechanism of preference reversal.

References and Notes

1. S. M. McClure, D. I. Laibson, G. Loewenstein, J. D. Cohen, *Science* **306**, 503 (2004).
2. G. Ainslie, *J. Exp. Anal. Behav.* **21**, 485 (1974).
3. M. Z. Deluty *et al*., *Behav. Anal. Lett.* **3**, 213 (1983).
4. D. Laibson, *Q. J. Econ.* **112**, 443 (1997).
5. L. Green, J. Myerson, *Psychol. Bull.* **130**, 769 (2004).
6. S. Frederick *et al*., *J. Econ. Lit.* **40**, 351 (2002).
7. G. Ainslie, J. Monterosso, in *Choice, Behavioural Economics and Addiction*, R. E. Vuchinich, N. Heather, Eds. (Elsevier, Oxford, 2003), pp. 35–61.
8. Gibbon, *Psychol. Rev.* **84**, 279 (1977).
9. K. Kirby, B. Guastello, *J. Exp. Psychol. Appl.* **7**, 154 (2001).
10. G. Ainslie, J. Monterosso, *J. Exp. Anal. Behav.* **79**, 37 (2003).
11. G. Ainslie, in *Disorders of Volition*, N. Sebanz, W. Prinz, Eds. (MIT Press, Cambridge, MA, in press).
12. G. S. Berns *et al*., *J. Neurosci.* **21**, 2793 (2001).
13. B. Knutson *et al*., *J. Neurosci.* **21**, RC159 (2001).
14. W. Schultz *et al*., *J. Neurosci.* **12**, 4595 (1992).
15. M. Mather *et al*., *Psychol. Sci.* **15**, 259 (2004).
16. M. P. Paulus *et al*., *Neuroimage* **13**, 91 (2001).
17. S. A. Bunge *et al*., *Neuroimage* **17**, 1562 (2002).
18. G. Loewenstein, *Organ. Behav. Hum. Decis. Process.* **65**, 272 (1996).
19. P. R. Montague, G. S. Berns, *Neuron* **36**, 265 (2002).
20. D. A. Hume, *A Treatise of Human Nature* (Oxford Univ. Press, Oxford, 1968; originally published 1739).
21. R. A. Rescorla, *Am. Psychol.* **43**, 151 (1988).
22. G. Ainslie, *Breakdown of Will* (Cambridge Univ. Press, Cambridge, 2001).

PHYSICS

The Environment Matters—Even on the Atomic Scale

Matthias Bode

A detailed understanding of magnetic excitations is essential for the future progress of magnetic data storage technologies. On page 466 of this issue, Heinrich *et al.* (*1*) use a scanning tunneling microscope (STM) to elucidate one such excitation, namely the spin-flip of individual magnetic atoms that are dispersed on a nonmagnetic matrix and exposed to an external magnetic field. Such excitations can degrade the performance of high-density memories. Extensions of the new method may allow other magnetic excitations to be studied.

When highly diluted magnetic atoms in a nonmagnetic host matrix are exposed to an external magnetic field B, the electron potentials of spin-up and spin-down atoms become slightly different. The energy required to overcome the resulting energy gap in a spin-flip process amounts to twice the Zeeman energy $E_Z = g\mu_B B$. Because B is an adjustable experimental parameter and the Bohr magneton μ_B is a fundamental constant, this relation can be used to measure the Landè g factor, which determines the spin and orbital contributions to the total magnetic moment.

Traditionally, the Zeeman energy is measured with electron spin resonance (ESR), which—due to sensitivity limitations—requires at least 10^7 electron spins. Therefore, a g value determined by ESR is averaged over a large number of supposedly identical magnetic atoms (*2*). However, the individual properties of the magnetic atoms may be rather different, because their local environment differs structurally, chemically, or both. Heinrich *et al.* (*1*) now use STM to determine the g values of individual Mn atoms on Al_2O_3 by measuring single-atom spin-flip processes.

How can spin flips and other inelastic processes be measured with an STM? The tunneling current between the tip and the sample is carried by elastic and inelastic "channels." In an elastic tunneling process, the energy of the electron is conserved when it hops out of an occupied state of the negatively biased electrode into an empty state of the positive one. In contrast, an inelastic tunneling process requires energy to be transferred between the tunneling electron and the sample. Because this energy is quantized, inelastic channels cannot contribute to the total tunneling current if the bias potential is lower than the quantization energy. Above this threshold, there will be a sudden conductance jump between tip and sample. This effect is the basis of inelastic scanning tunneling spectroscopy (STS), which has provided profound insights into vibrational resonances of single molecules (*3*).

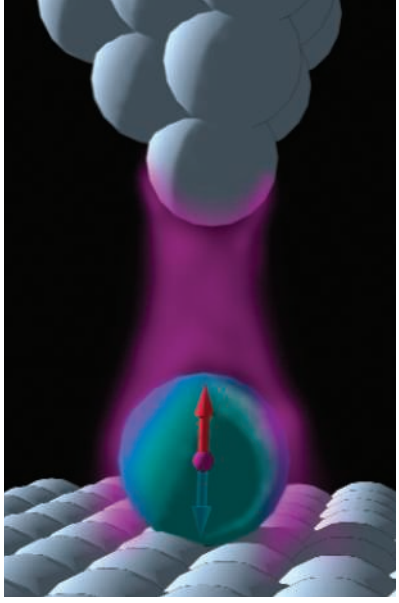
The method used by Heinrich *et al.* is an extension of inelastic STS. The authors exploit the fact that magnetic excitations, such as spin-flip excitations, are sensitive to an external magnetic field. The threshold energy for magnetic excitations increases with increasing field strength, but this effect is extremely small (typically <1 meV/tesla). To measure such tiny energy shifts with sufficient sensitivity, Heinrich *et al.* used a home-built STM, which operates at 0.6 K (reducing thermal broadening) and is mounted inside a superconducting magnet that supplies up to 7 tesla. They find that the g value of Mn atoms depends on their adsorption site: A Mn atom close to an Al_2O_3 step edge has a higher g value than a Mn atom far away from a step edge.

A detailed understanding of magnetic excitations is not only of academic interest, but is essential for future increases in the storage density of magnetic memories. The exponential increase in storage density achieved over the past 50 years (*4*) was mainly based

The author is with the Institute of Applied Physics and Microstructure Research Center, University of Hamburg, Jungiusstrasse 11, Hamburg 20355, Germany. E-mail: mbode@physnet.uni-hamburg.de

on shrinking of classical devices without changing the basic technology. With the lateral dimensions of magnetic bits falling below 10 nm, however, the atomic structure of the materials and their interfaces begins to have detrimental effects on device performance. Therefore, further increases in storage density will require new concepts.

One such new concept is the magnetic random access memory (MRAM), which uses the spin of tunneling electrons rather than the electron charge to store and read information (*5*). For proper function of the MRAM, the spin of a tunneling electron has to be conserved because it carries the information. But impurities embedded in the tunneling barrier or at the interfaces of the MRAM lead to unwanted spin-flip processes. Up to now, our understanding of these processes was hampered by the poor characterization of the chemical nature of the impurities and of their environment [see (*1–9*) in (*1*)]. As shown by Heinrich *et al*., the STM's ability to perform atomic-scale imaging and spectroscopy—especially when combined with single-atom manipulation—removes many ambiguities and may eventually help to avoid spin-flip scattering in MRAM.



An artist's view of single-atom spin-flip excitations induced by means of STM. If the energy of the electrons tunneling between tip and sample is sufficiently large, spin-flip processes occur, leading to a sudden conductance jump. Heinrich *et al.* (*1*) show that the threshold energy for spin-flip processes depends on the environment of magnetic atoms, that is, the coordination of adjacent atoms and their chemical composition.

Self-organization of nanometer-scale magnetic particles (*6*) is another concept proposed for high-density data storage. The time required to write information into a particular particle (bit) strongly depends on the creation and damping of collective magnetic excitations, called magnons, which affect atomic spins and can be envisioned as small-amplitude oscillations propagating through the particle (*7*). The role of atomic-scale defects in the creation and propagation of magnons is largely unexplored, mainly due to the lack of suitable measurement techniques. Inelastic STS may close this gap.

Heinrich *et al*. have shown that atomic-scale magnetic excitations can be localized with inelastic STM. But their measurement technique is still somewhat indirect, because the initial and final spin states of the magnetic atoms (spin-up or spin-down) were not visualized. Ultimately, by using magnetic tips, inelastic STS may be combined with the ability to image the magnetization direction directly, as accomplished with so-called spin-polarized STM (*8–10*), leading to further insights into current problems of atomic-scale magnetism.

For example, the method may be used to measure correlation effects between the delocalized electrons of nonmagnetic matrices and unpaired electrons of magnetic impurities (the Kondo effect). Non–spin-resolved STS has revealed a strong dependence of the Kondo effect on the impurity size (monomer, dimer, or trimer), which was attributed to differences in the spin configurations (*11*), but a definite proof is still lacking. Another interesting question is whether single atoms can be magnetically stable and, if so, how they can be switched. The recently discovered huge anisotropy of Co atoms on the Pt(111) surface (*12*) indicates that—at sufficiently low temperature—they may indeed be the smallest possible permanent magnets, but nobody has imaged them yet.

References and Notes

1. A. J. Heinrich, J. A. Gupta, C. P. Lutz, D. M. Eigler, *Science* **306**, 466 (2004); published online 9 September 2004 (10.1126/science.1101077).
2. Rugar *et al.* have recently described a method for detecting the spin of a single electron with magnetic resonance force microscopy (*13*).
3. W. Ho, *J. Chem. Phys.* **117**, 11033 (2002).
4. E. Grochowski, R. D. Halem, *IBM Syst. J.* **42**, 338 (2003)
5. A. Cho, *Science* **296**, 246 (2002).
6. S. Sun *et al.*, *Science* **287**, 1989 (2000).
7. B. Hillebrands, K. Ounadjela, Eds., *Spindynamics in Confined Magnetic Structures II* (Springer, Berlin, 2003).
8. M. Bode, *Rep. Prog. Phys.* **66**, 523 (2003).
9. S. Heinze *et al.*, *Science* **288**, 1805 (2000).
10. R. Wiesendanger, *Curr. Opin. Solid State Matter Sci.* **4**, 435 (1999).
11. T. Jamneala, V. Madhaven, M. F. Crommie, *Phys. Rev. Lett.* **87**, 256804 (2001).
12. P. Gambardella *et al.*, *Science* **300**, 1130 (2003).
13. D. Rugar *et al.*, *Nature* **430,** 329 (2004).

BIOMEDICINE

Insulin Resistance Takes a Trip Through the ER

Deborah M. Muoio and Christopher B. Newgard

Type 2 diabetes, which afflicts about 150 million people worldwide, has emerged as one of the leading global health threats of the 21st century (*1*). Diabetes develops when resistance to the glucose-lowering actions of insulin combines with impaired insulin secretion, giving rise to dangerously high concentrations of glucose in the blood. The prediabetic onset of insulin resistance is usually preceded by weight gain—more than 80% of type 2 diabetics are overweight (*2*). Given that experts are forecasting little reprieve from the current obesity epidemic, efforts to understand the molecular mechanisms that connect the two diseases have intensified. On page 457 of this issue, Özcan *et al.* (*3*) propose that the protein production factory of mammalian cells, the endoplasmic reticulum (ER), is an important sensor of metabolic stress that may render insulin powerless to maintain systemic glucose homeostasis.

The authors are in the Sarah W. Stedman Nutrition and Metabolism Center, and Departments of Pharmacology and Cancer Biology, Medicine, and Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. E-mail: newga002@mc.duke.edu

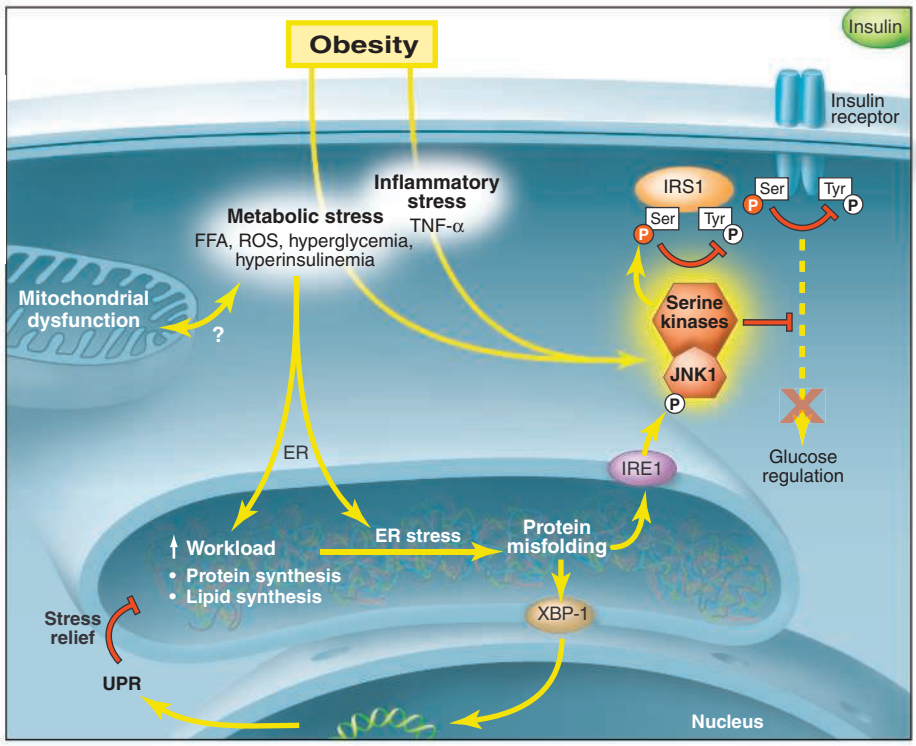
The signaling pathways that mediate insulin action depend on a tyrosine-phosphorylation cascade that begins with autoactivation of the insulin receptor tyrosine kinase, followed by tyrosine phosphorylation of proximal targets such as insulin receptor substrate 1 (IRS1) (*4*). Some forms of insulin resistance may be mediated by a serine kinase cascade that targets the insulin receptor or its downstream signaling partners (see the figure) (*4–7*). In contrast to tyrosine phosphorylation, which serves to propagate the signal, serine phosphorylation prevents the signal from reaching its final destination. Earlier research identified the c-Jun amino-terminal kinase (JNK1) as a new member of the growing network of serine kinases that inhibit insulin signaling (*5*). JNK is a central regulator of inflammatory and immune responses, but its role in metabolic control is less certain. Like other members of this network, JNK1 is activated by free fatty acids and the inflammatory cytokine, tumor

Published by AAAS

necrosis factor–α (TNF-α). The subsequent discovery that obesity increases JNK activity in insulin-responsive tissues such as fat, muscle, and liver pointed to a potential link between inflammatory and insulin-desensitizing signaling pathways. Consistent with this idea, targeted deletion of JNK1 in mice prevented obesity-induced serine phosphorylation of hepatic IRS1, and likewise protected animals from insulin resistance.

In an attempt to better understand the obesity-associated events linked to JNK1, Özcan *et al*. discovered that activation of this kinase can be triggered by ER stress. The ER is a membranous network that provides a specialized environment for processing and folding newly synthesized proteins (see the figure). Thus, as metabolic demands increase, so too does the workload of this protein factory. Biological insults such as infection, hypoxia, nutrient deprivation, and exposure to chemical toxins or excess lipids can disrupt ER homeostasis, causing unfolded or misfolded proteins to accumulate in the ER lumen (*8*). To alleviate this stress, the ER initiates a transcriptional program referred to as the unfolded protein response, which slows protein synthesis and promotes protein degradation (*8*).

Özcan *et al*. postulated that obesity might impose a strain on the ER machinery, thereby triggering a response that activates JNK1 and impairs the insulin signaling pathway (*3*). They found that markers of ER stress, along with activated JNK1, were indeed elevated in the adipose tissue and liver of mice with genetic or diet-induced forms of obesity. In cultured liver cells, pharmacologically induced ER stress caused increased JNK activity, IRS1 serine phosphorylation, and impaired insulin signaling, whereas treatment with a JNK inhibitor blocked these stress-induced events. Further support for their hypothesis came from genetic models in which the ER stress response was modified by altered expression of a proximal ER stress-sensor called inositol requiring kinase 1 (IRE1), or of XBP-1, a downstream transcription factor that modulates the unfolded protein response. In embryonic fibroblasts from IRE1-deficient mice, chemical ER stressors were unable to activate JNK1; consequently, the cells were protected against insulin resistance. Similar protective effects were observed in cultured liver cells that overexpressed XBP-1 and hence were better prepared to cope with ER stress. In these cells, an increase in XBP-1 prevented JNK1 activation in response to the chemical insult, presumably due to enhancement of the unfolded protein response. Moreover, heterozygous deletion of XBP-1 in a strain of mice normally resistant to diet-induced diabetes produced mice that were prone to the disease. The increased susceptibility of the XBP-1 heterozygous mice was associated with chronic ER stress, JNK1 hyperactivation, and impaired insulin signaling in the liver.



No stress relief for the ER. The metabolic and inflammatory stresses of obesity disrupt the smooth operation of the ER and cause protein misfolding. The ER attempts to cope with stress by activating XBP-1, a transcriptional regulator of the unfolded protein response (UPR). If these responses fail to restore homeostasis, stress-induced IRE1 activates JNK1, a serine kinase that opposes insulin action. Impaired insulin signaling might serve to alleviate intracellular stress, but it does so at the expense of systemic glucose regulation. FFA, free fatty acids; ROS, reactive oxygen species.

Obesity is associated with metabolic and inflammatory stresses that combine to mount a full-scale systemic attack on glucose homeostasis. Özcan *et al*. add a new element to this picture. They portray obesity as a state in which molecular signals launched by a distressed ER contribute to impaired insulin action. Whether obesity-induced disturbances in the ER stem from chronic lipid overload, the anabolic pressures of hyperinsulinemia, cytokine-induced signaling, mitochondrial dysfunction, or other pathophysiological assaults now awaits further investigation. Intriguingly, the enzymes responsible for processing excess lipid include several integral membrane proteins that reside in the ER.

The Özcan *et al*. findings also question the extent to which ER stress might explain the tissue disturbances associated with diabetes. Indeed, interruption of a signaling event involved in relieving ER stress—the phosphorylation of eukaryotic translation initiation factor–2 by pancreatic ER kinase (PERK)—results in severe functional impairment of pancreatic islet β cells (*9*, *10*). However, feeding rodents a high-fat diet, which causes insulin resistance in liver and muscle in most rodent strains, is not sufficient to impair insulin secretion in islet β cells. Moreover, Özcan and colleagues did not find evidence of ER stress and the unfolded protein response in skeletal muscle, even though this tissue is thought to be responsible for most systemic glucose disposal. Does this mean that muscle and islet β cells are less susceptible to ER stress and the unfolded protein response associated with overnutrition or other metabolic stressors? Alternatively, do signals generated in liver and adipose tissue in response to ER stress and the unfolded protein response contribute to the ultimate failure of pancreatic islets and muscle to secrete and sense insulin, respectively? As scientists such as Özcan and co-workers contribute new molecular clues, opportunities for therapeutic advancement continue to expand for the patients who battle the ravages of these diseases.

References

1. P. Zimmet *et al*., *Nature* **414**, 782 (2001).
2. M. M. Engelgau *et al*., *Ann. Intern. Med.* **140**, 945 (2004).
3. U. Özcan *et al*., *Science* **306**, 457 (2004).
4. A. R. Saltiel, J. E. Pessin, *Trends Cell Biol.* **12**, 65 (2002).
5. J. Hirosumi *et al*., *Nature* **420**, 333 (2002).
6. G. Perseghin *et al*., *Int. J. Obes. Relat. Metab. Disord.* **27** (suppl. 3), S6 (2003).
7. S. E. Shoelson *et al*., *Int. J. Obes. Relat. Metab. Disord.* **27** (suppl. 3), S49 (2003).
8. E. Szegezdi *et al*., *Ann. N.Y. Acad. Sci.* **1010**, 186 (2003).
9. H. P. Harding *et al*., *Mol. Cell* **7**, 1153 (2001).
10. D. Scheuner *et al*., *Mol. Cell* **7**, 1165 (2001).

CREDIT: KATHARINE SUTLIFF/*SCIENCE*

EPPENDORF ESSAY WINNER

Deconstructing *C. elegans* Sensory Mechanotransduction

Miriam B. Goodman

Eppendorf and *Science* are pleased to present the prize-winning essay by Miriam Goodman, the 2004 winner of the Eppendorf and *Science* prize for Neurobiology

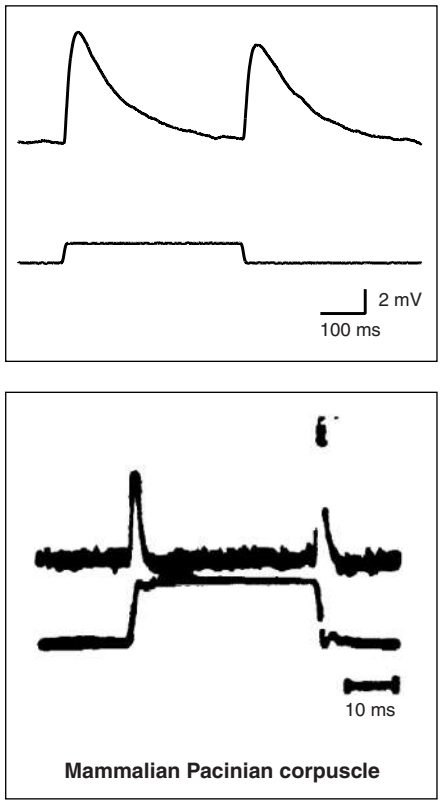
eppendorf & Science PRIZE FOR NEUROBIOLOGY

We as human beings depend greatly on the sensory neurons that govern our sense of touch. If such cells cease to function properly, we may lose the ability to respond to another's touch—say, as a dancer responds to a partner's lead. Additionally, we may be unable to respond to a more painful and potentially damaging event. Reduced touch sensation, or peripheral sensory neuropathy, is especially common in people with diabetes and is a significant contributing factor to lower-extremity amputations (*1*). Other mechanoreceptor neurons mediate equally vital sensory modalities: proprioception for balance and the control of internal organs, such as the bladder and kidney, and baroreception for homeostatic control of heart rate. Despite the importance of these mechanical senses, however, exactly how sensory cells detect the mechanical energy in a touch, the bend of a limb, or changes in blood vessel diameter remains a mystery.

Although electrical responses to mechanical stimuli were first measured in the 1920s (*2*, *3*), surprisingly little is known about the protein machinery that converts mechanical energy into ionic currents in touch-sensitive neurons, and even less is known about how the individual protein components of this nanomachine operate. Research into the molecular basis of touch transduction lags behind research into other senses because, in many animals, sensory nerves that detect touch are scattered across the body and are deeply embedded in the skin—two properties that complicate traditional biochemical approaches. To circumvent these difficulties, we study touch sensation in the nematode worm *Caenorhabditis elegans*. This is the only animal for which we know the cellular anatomy of the entire nervous system. Compared with tens of thousands cutaneous sensory neurons in mammals, each worm has only six nerve cells that govern touch sensation along its body wall (*4*). Genetic analyses by Martin Chalfie and his colleagues revealed that the worm's sense of touch requires at least 12 specific proteins, encoded by the *mec* or mechanosensory abnormal genes [reviewed in (*5*)].

To understand how proteins identified by genetic screens contribute to mechanotransduction, my collaborators and I developed methods for in vivo recording from identified sensory neurons in *C. elegans* (*6*) and used these methods to record electrical responses to external force in *C. elegans* touch neurons (*7*). The initial experiments focused on a pair of neurons that sense gentle touch to the worm's tail. We found that receptor potentials in *C. elegans* touch neurons are reminiscent of the responses of vibration-sensitive Pacinian corpuscles measured 40 years ago in mammals (*8*), which suggests that aspects of mechanotransduction may be similar in nematodes and mammals (see the figure). Activation of mechanoreceptor currents (MRCs) in *C. elegans* touch neurons is extremely rapid: Current begins to flow within 1 ms of force application. It is the first step in transduction, preceding both membrane depolarization (*7*) and increases in somatic Ca^{2+} (*9*). Such latencies are nearly two orders of magnitude faster than those reported for *Drosophila* phototransduction (20 to 100 ms), the current record-holder for a second messenger–mediated G protein signaling cascade (*10*). Thus, external force might open ion channels directly rather than operating via a separate force receptor.

We are also working to deconstruct the mechanotransduction complex by asking how mutations in *mec* genes alter MRCs in vivo. The first mutations we studied eliminate or alter four membrane proteins (MEC-4, MEC-10, MEC-2, and MEC-6), which are believed to form the ion channel at the core of the transduction complex. Consistent with this idea, null mutations in *mec-4*, *mec-2*, and *mec-6* abolish MRCs without affecting other ion currents, which indicates that these proteins are specifically required for the generation of MRCs and are likely to encode subunits of the ion channels that carry MRCs in vivo. Additionally, an allele of *mec-10*, which substitutes glutamate for a conserved glycine residue near the second transmembrane domain of MEC-10 (*11*), reduced but did not eliminate MRCs. This reduction appears to result from altered ion selectivity, as opposed to a genetic deletion of transduction channels. In short, all mutations that diminish touch sensation abolish or alter MRCs in vivo. Our findings link the application of external force to the activation of a molecularly defined sensory transduction channel.



Responding to touch. Comparision of mechanoreptor potentials in *C. elegans* touch receptor neurons (**top**) and mammalian Pacinian corpuscles (**bottom**). [Data in the bottom panel are adapted from (*8*)]

The author is in the Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA 94305, USA. E-mail: mbgoodman@stanford.edu

Published by AAAS

In addition to analyzing electrical responses to external force in vivo, we have taken the first steps toward reconstituting this ion channel complex in *Xenopus* oocytes. So far, we have bypassed the need for mechanical gating by studying a constitutively active mutant channel (the "d" form). Coexpression of MEC-4d and MEC-10d produces a constitutively active current that, like native MRCs (*7*), is carried by Na^+ and blocked by amiloride (*12*). By contrast with native MRCs, however, neither MEC-2 nor MEC-6 was required to produce detectable channel activity in oocytes (*12*, *13*). Both accessory proteins increased activity of expressed MEC-4d/10d channels at least tenfold without inducing a detectable increase in surface expression of either MEC-4 or MEC-10 (*12*, *13*), which suggests that MEC-2 and MEC-6 increase single-channel conductance or open probability. Preliminary studies of single MEC-4d/10d channels suggest that neither MEC-2 nor MEC-6 significantly increases single-channel conductance, however (*14*). Additional studies of both expressed and native channels are needed to clarify the function of these accessory proteins in mechanotransduction.

By demonstrating that native MRCs require intact copies of *mec-4*, *mec-10*, *mec-2*, and *mec-6*, we show that each of these four genes is required for the first step in mechanotransduction—namely, activation of sensory mechanotransduction channels. Such channels may be directly gated by mechanical energy, because MRCs can be detected within 1 ms of stimulation. Because both *C. elegans* touch receptor neurons and mammalian Pacinian corpuscles respond preferentially to changes in force (*7*, *8*), we speculate that DEG/ENaC channels could be sensory mechanotransduction channels in nonciliated mechanoreceptor neurons in nematodes and mammals alike. These initial studies raise new questions, such as: How do touch receptor neurons detect changes in force while remaining insensitive to continuous force application? How is force transferred from the worm's cuticle to transduction channels? What determines sensitivity? A better understanding of the answer to this last question could lead to improved diagnosis and treatment of sensory neuropathy.

References

1. Centers for Disease Control and Prevention, "National diabetes fact sheet: General information and national estimates on diabetes in the United States" (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2004).
2. E. D. Adrian, *J. Physiol.* **61**, 49 (1926).
3. E. D. Adrian, Y. Zotterman, *J. Physiol.* **61**, 151 (1926).
4. M. Chalfie *et al.*, *J. Neurosci.* **5**, 956 (1985).
5. G. G. Ernstrom, M. Chalfie, *Annu. Rev. Genet.* **36**, 411 (2002).
6. M. B. Goodman, D. H. Hall, L. Avery, S. R. Lockery, *Neuron* **20**, 763 (1998).
7. R. O'Hagan, M. Chalfie, M. B. Goodman, submitted.
8. M. Mendelson, W. Loewenstein, *Science* **144**, 554 (1964).
9. H. Suzuki *et al.*, *Neuron* **39**, 1005 (2003).
10. R. C. Hardie, *J. Exp. Biol.* **204**, 3403 (2001).
11. M. Huang, M. Chalfie, *Nature* **367**, 467 (1994).
12. M. B. Goodman *et al.*, *Nature* **415**, 1039 (2002).
13. D. S. Chelur *et al.*, *Nature* **420**, 669 (2002).
14. A. L. Brown, M. B. Goodman, unpublished data.

15. The research described here is the happy result of excellent collaborations with scientists at the University of Oregon, Columbia University, and Stanford University. It would not have been possible except by working jointly. I thank all of you. Research in my lab is supported by fellowships from the Alfred P. Sloan Foundation, the Donald B. and Delia E. Baxter Foundation, and a grant from the National Institute of Neurological Disorders and Stroke.

2004 Grand Prize Winner

Dr. Miriam B. Goodman grew up in Lexington, Massachusetts, and Bethesda, Maryland. As a high school student, she worked in research labs at the National Institutes of Health where she wrote scientific software. She earned a bachelor's degree in Biochemistry from Brown University in 1986. As a graduate student in neurobiology at the University of Chicago, she analyzed voltage-dependent ion channels that tune vertebrate hair cells. After being awarded her Ph.D. in 1995, she pursued postdoctoral work in *C. elegans* neurophysiology and genetics at the University of Oregon and Columbia University. Currently, Dr. Goodman is an Assistant Professor of Molecular and Cellular Physiology at Stanford University. Work in her laboratory focuses on delineating the molecular events that give rise to the sense of touch. Outside the laboratory, Dr. Goodman enjoys cooking with friends, hiking, rock-climbing, and going to the movies. Though currently sidelined, Dr. Goodman has also played soccer since age 8.

Finalists

Kang Shen, for his essay "Synaptic Matchmakers: Molecular Mechanisms of Synaptic Specificity." Dr. Shen was born and raised in Wuhan, China. He studied clinical medicine at Tongji Medical University of China. After graduating in 1994, he joined the graduate program at Duke University, where he studied the spatial and temporal control of CaMKII localization in hippocampal neurons in the laboratory of Dr. Tobias Meyer. After receiving his Ph.D. in 1999, he pursued postdoctoral work in Dr. Cornelia Bargmann's lab at the University of California, San Francisco, where he addressed the question of synaptic specificity, using *C. elegans* as a model system. Dr. Shen started his own lab at Stanford University in 2003, focusing on understanding molecular mechanisms of synaptic target specificity. Outside of the laboratory, Dr. Shen enjoys a variety of sports and outdoor activities.

Qin Shen, for her essay "Preventing Aging in Neural Stem Cells: Regulating Asymmetric Versus Symmetric Cell Divisions." Dr. Shen was born and grew up in China. She earned her Bachelor's degree in Pharmacology from Shanghai Medical University in 1991. In 1996, she entered the graduate program in Neuroscience at Albany Medical College, New York, under the guidance of Dr. Sally Temple, who specializes in neural stem cell development. Her Ph.D. project, completed in 2001, focused on asymmetric cell division and the generation of cell diversity in the embryonic murine cerebral cortex. She is now a postdoctoral fellow in Dr. Temple's laboratory working on mechanisms regulating neural stem cell self-renewal and cell fate choices, including interactions between neural stem cells and endothelial niche cells. The mother of a toddler, Dr. Shen also carves out a little time for gardening and reading.

For the full text of essays by the finalists and for information about applying for next year's awards, see *Science* Online at http://www.sciencemag.org/feature/data/prizes/eppendorf/eppenprize.shtml

INTRODUCTION

Neuroscience: Higher Brain Functions

What are the brain substrates activated during higher cortical functions, such as cognition, and how do they interact with structures involved in guiding our behavior? Until recently, these questions were discussed completely independently in the faculties of neurobiology and psychology, and on a different level in economics and linguistics. However, in the past couple of years we have witnessed an enormous growth in interdisciplinary studies, in which some of the leading experts in their fields have branched out of the traditional confines of their specialities and achieved invaluable new insights into the biological basis of thought processes traditionally considered to belong more to the realm of the humanities. This has led to the rapid expansion of new fields, such as neuroeconomics. In this year's neuroscience special issue, we have tried to give a synopsis of recent developments in this successful example of scientific cross-fertilization.

Miyashita's Review on cognitive memory (p. 435) attempts to integrate molecular, cellular, electrophysiological, and functional imaging data dealing with questions of encoding and retrieval of episodic memory. Ridderinkhof *et al*. (p. 443) review the literature and perform a meta-analysis of the role of the posterior medial frontal cortex in functions of cognitive control, such as monitoring of unfavorable outcomes, response errors, and response conflict. They suggest that the posterior medial frontal cortex is commonly activated when the need for performance adjustment becomes evident to the individual. Glimcher and Rustichini (p. 447) describe the latest developments in neuroeconomics, which tries to combine modern neurobiological techniques with well-established traditional approaches in economics and psychology. In a Viewpoint, Gelman and Gallistel (p. 441) discuss two papers in this issue of *Science* by Pica *et al*. (p. 499) and by Gordon (p. 496) on exciting new findings concerning the role of language in the origin of numerical concepts. They discuss these findings in the broader context originally proposed by Benjamin Lee Whorf: how language not only influences but constrains the way we experience the world and even how we think.

CREDIT: BURSTEIN COLLECTION/CORBIS

In a News story, Miller (p. 432) visits Denmark for the latest International Congress of Neuroethology and reveals how this field uses the remarkable diversity of the animal kingdom to understand the ways in which a brain controls behavior. Possible connections between behavior, personality, and aging in both humans and monkeys are discussed by Corr in a Perspective in the Science of Aging Knowledge Environment (SAGE KE, www.sciencemag.org/sciext/cognition).

In *Science*'s Signal Transduction Knowledge Environment (STKE, www.sciencemag.org/sciext/cognition), Cull-Candy and Leszkiewicz review the discrete roles played by various NMDA receptor subtypes. Pollack discusses a signaling pathway mediated by coactivated dopamine D1 and D2 receptors that is distinct from those activated by D1 or D2 alone. An animation by Contractor and Heinemann depicts AMPA receptor cycling at the synapse.

–PETER STERN, GILBERT CHIN, AND JOHN TRAVIS

CONTENTS

See also related STKE and SAGE KE material on p. 367, and Editorial on p. 373.

Science

Published by AAAS

NEWS

Behavioral Neuroscience Uncaged

Aided by dancing spiders, electric fish, and assorted frogs, neuroethologists continue a proud tradition of raiding nature's menagerie for insights into brain and behavior

NYBORG, DENMARK—All eyes were on the giant spider. Projected larger than life onto the screen at the front of the darkened auditorium, a sturdy specimen of the jumping spider *Habronattus dossenus* stared back with at least four of its eyes and fidgeted its hairy legs. Then it began to dance.

Male jumping spiders are renowned for their elaborate courtship displays, but Ronald Hoy, a neuroscientist at Cornell University in Ithaca, New York, was showing off a talent previously unknown for this arachnid. Aided by an amplified audio feed, Hoy demonstrated that the spider creates an auditory and seismic accompaniment to its visual display. As the spider strutted to and fro, scraping sounds and buzzes accentuated its movements, which were punctuated occasionally by a thump—the spider's abdomen slapping the ground—and a dramatic outward thrust of two front legs. Hoy told the audience that the spider's rhythm reminds him of flamenco, and a video comparison of the two dance forms brought murmurs of agreement.

Hoy's graduate student Damian Elias has found that the sound and seismic signals made by male jumping spiders are key to their mating success. Elias now plans to investigate vibration sensors in the spiders' legs that enable them to catch each other's vibes.

Hoy's presentation at the 7th International Congress of Neuroethology* illustrated a belief held passionately by the more than 500 researchers in attendance: Animal behavior—in particular, that of animals striving to survive and multiply in their natural environments—is fascinating to behold. Studying such behaviors and their neural underpinnings is the heart and soul of neuroethology.

At the congress, scientists from 19 countries presented work on a wide range of topics that have proven amenable to the neuroethological approach. The presentations included new findings on how the brain deciphers information gathered by the senses and also about the neural mechanisms of communication and movement control. Overall, the meeting's talks and posters, which featured a menagerie of creatures beyond the standard lab animals, illustrated how for more than half a century, neuroethologists have mined the tremendous diversity of behaviors and nervous systems in nature for general principles about how brains—including ours—direct actions.

Neuro-what?

Neuroethology is generally not well known as a field, concedes Edward Kravitz, a researcher at Harvard University and incoming president of the International Society for Neuroethology, which organizes the triennial congress. The field traces its roots to pioneering studies in the first half of the last century by the likes of Karl von Frisch, Konrad Lorenz, and Nikolaas Tinbergen. Their work on the natural behavior of bees, birds, and fish highlighted the importance of interpreting animal behavior in the context of its survival value and showed that much could be gained by comparing behaviors across species. The fruits of this approach, which became known as ethology, earned the three a Nobel Prize in 1973.

¡Olé! The dance moves of male jumping spiders don't just look good. They produce seismic signals that put females in the mood.

CREDIT: DAMIAN ELIAS

Neuroethology goes one step beyond ethology to ask how the nervous system controls behavior. Its guiding principles include a bit of wisdom passed on by the Danish physiologist August Krogh, who suggested that for any question a biologist might care to pose, there is a species particularly well suited to provide an answer.

One way neuroethologists apply Krogh's principle is by studying species with exceptional talents. Owls are an oft-cited example. The birds hunt at night by localizing the rustles and squeaks made by their prey. By probing the auditory regions of owl brains, researchers have learned how the brain creates a spatial map from cues such as differences in the time it takes a sound to reach the two ears. This knowledge has clarified how less-expert species like ourselves localize sounds and, more generally, has illuminated how brains accomplish the seemingly impossible task of tracking events 1000 times more fleeting than a single nerve impulse.

"It's the specialist species more than the generalist species that tell us what biology can do," says Hermann Wagner, who studies barn owls at the University of Aachen in Germany. When it comes to sound localization, he adds, "the barn owl is a Ferrari. … If you want to understand engines, wouldn't you rather study a Ferrari?"

Another application of Krogh's principle has led researchers to creatures with relatively simple nervous systems, which often happen to have exceptionally large—and therefore easily accessible—neurons. Studies of sea slugs, for example, have provided Nobel Prize–winning insights into how learning and memory modify the chemical communication pathways between and within neurons.

The second guiding principle of neuroethology is the comparative approach of the early ethologists. By comparing profiles of gene expression in the brains of songbirds with those of birds that don't sing, for example, researchers have begun to uncover clues about the evolution of vocal communication.

A bird's-eye view

When Merlin the magician oversees the education of the young King Arthur in T. H. White's *The Once and Future King*, he turns Arthur into a fish, an ant, and various birds so that the young king will experience a variety of perspectives on the world. As that legend suggests, the sensory realm of animals is extraordinarily rich,

*8–13 August, Nyborg, Denmark.

Published by AAAS

and it has provided neuroethologists with fertile ground for study.

From a human point of view, one of the more foreign ways of taking the measure of the world is the sense of magnetoreception used by migratory birds to navigate between their winter homes and summer breeding grounds. The biological mechanism for the birds' magnetic compass has been a contentious topic for decades. In a presentation that generated animated coffee-break discussion at the congress, Henrik Mouritsen of the University of Oldenburg in Germany offered new experimental support for a rather startling hypothesis: that migrating birds literally see Earth's magnetic field.

A team from the University of Illinois, Urbana-Champaign, first raised that possibility in a theoretical paper published in the *Biophysical Journal* in 2000. It suggested that light-sensitive molecules called cryptochromes could form the basis of a biological compass.

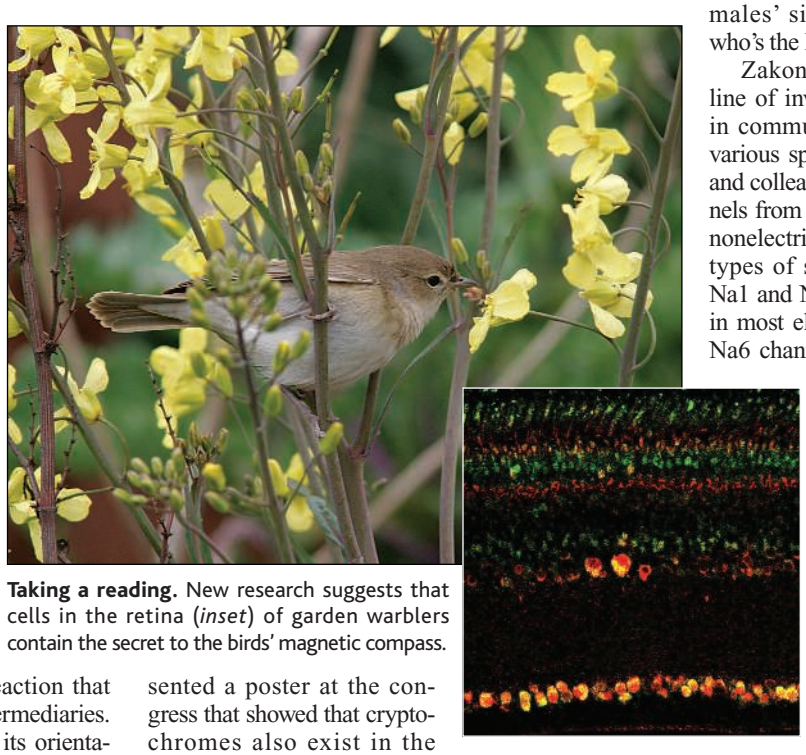
The general idea is this: In response to light, cryptochromes undergo a chemical reaction that creates so-called radical-pair intermediaries. A magnetic field, depending on its orientation, can alter the spin state of electrons in the radical pair and tweak the ratio of two final reaction products, which somehow turns the dial up or down on the chemical cascades that normally convert light to nerve impulses in the retina.

The implication, Mouritsen says, is that if cells in the bird retina contained cryptochromes, the planet's magnetic field would modulate the cells' sensitivity differently in different parts of the retina. "It might be that a bird perceives a ghost image of the magnetic field on top of whatever else it sees."

Using fluorescently tagged antibodies, Mouritsen's team found that cryptochromes are expressed in certain cells in the retinas of migratory garden warblers but not in nonmigratory zebra finches. The team also found that in the warblers, but not the finches, genes related to neural activity fire up in these cryptochrome-carrying neurons during magnetic orientation in the early evening—the time of day when wild birds take bearings from their magnetic compass (*Science*, 16 April, p. 405). Mouritsen recently published the cryptochrome findings in the *Proceedings of the National Academy of Sciences*.

The results look promising, says Wagner. "It's indirect evidence, and I think much more needs to be done, but it's the first solid clue that cryptochromes play a role in magnetic orientation," he says.

And Mouritsen isn't the only one hunting cryptochromes. Andrea Möller, a Ph.D. student at Johann Wolfgang Goethe University in Frankfurt am Main, Germany, presented a poster at the congress that showed that cryptochromes also exist in the retinas of migratory robins.



Taking a reading. New research suggests that cells in the retina (*inset*) of garden warblers contain the secret to the birds' magnetic compass.

CREDITS (TOP TO BOTTOM): H. MOURITSEN; (INSET) H. MOURITSEN *ET AL.*, *PNAS*, EARLY ONLINE EDITION (20 SEPTEMBER 2004)

Lines of communication

The way an animal perceives the world can determine the way it keeps in touch with its comrades. Consider electric fish, another favorite research subject of neuroethologists. The fish navigate through murky waters by creating an electric field and monitoring the field for distortions caused by obstacles. They generate the field by producing pulses with their electric organ, a modified muscle in the tail.

To navigate successfully, a fish has to keep track of its own electric-organ discharges and ignore those of its neighbors. Studying how they do this has revealed general principles about how animals sort through a barrage of conflicting sensory information and home in on cues that matter.

The fish also use their electric-organ discharges for communication. Harold Zakon of the University of Texas, Austin, has been investigating how and why these signals differ from fish to fish. His team has found that in some electric fish species, male sex hormones suppress the expression of genes for a particular component of the sodium channels essential for the electric organ's discharge. Channels built from these components turn on very quickly, enabling a rapid string of short pulses. Suppressing the genes for these subunits makes the males' pulses longer and reduces their frequency compared to those of females, an effect that's even more pronounced in the androgen-drunk males at the top of the social hierarchy. Variations in the males' signals may tell females who's the Big Kahuna, Zakon says.

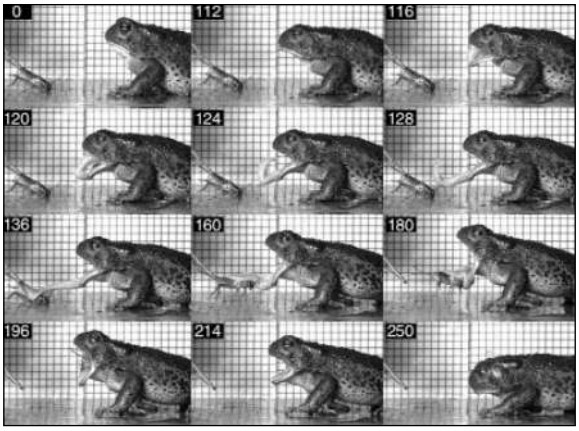
Zakon has now extended this line of investigation to differences in communication signals among various species of electric fish. He and colleagues cloned sodium channels from 11 species of electric and nonelectric fish. All fish have two types of sodium channels, called Na1 and Na6, in their muscles. But in most electric fish, Zakon found, Na6 channels have been lost from muscle and are only expressed in the electric organ. And unlike the highly conserved sodium channels found in nerve and muscle, the Na6 channels in the electric organ appear to vary from one electric fish species to the next. The researchers identified several differences in the amino acid sequence of the Na6 channels that they suspect alter the way the channels open and close and account for the wide variation in communication signals seen in different species.

Stick your tongue out, please

Keen perception and meaningful chatter are useless without the right moves to back them up. Understanding the neural control of movement is another major area of neuroethology research. Much of this work focuses on invertebrates. Crabs, slugs, leeches, and locusts have taught researchers a great deal about how circuits of neurons produce rhythmic behaviors such as walking, swimming, and flying. For many of these creatures, researchers have identified every neuron involved in producing a particular behavior, mapped out how one connects to another, and deciphered what chemicals they use to communicate—a level of understanding nearly impossible to achieve in more complicated nervous systems.

Such studies have clarified how sequences of electrical impulses from neurons generate the series of muscle contractions needed to bring about an intended motion. Two presentations at the congress, however, highlighted another aspect of controlling movement: the role of biomechanics.

Kiisa Nishikawa of Northern Arizona University in Flagstaff brought her colleagues up to date on her lab's work on how frogs zap tasty insects with their tongues.



Quick on the draw. Ramping up the tension in two jaw muscles primes a frog's tongue to snag a snack.

Over the last 20 years, she and her co-workers have studied nearly three dozen frog species and have identified three mechanisms of tongue projection. Their work has elucidated the evolutionary history of each mechanism as well as the interplay of nerve, muscle, and physics underlying all three.

In terms of brute force, the most impressive of the three mechanisms is called inertial elongation, which launches the tongue out of the mouth at 30 or more times the acceleration of gravity. The feat stretches the tongue to twice its resting length, substantially increasing the frog's range. This mechanism has evolved independently at least a half-dozen times, Nishikawa has found.

Coordinating this behavior would seem to be a complex task for the nervous system. It involves very fast, tightly coordinated movements of multiple muscles and joints. Yet Nishikawa's research has shown that much of the coordination that's required is built into the anatomy of the jaw and the physiology of its muscles.

Her modeling studies have revealed that more than 90% of the force for tongue projection comes from momentum transferred to the tongue from the lower jaw as it snaps open. The energy needed to launch the tongue is stored as elastic strain in a single pair of jaw muscles, and the anatomy of the jaw and tongue ensures that the tongue is flung forward on a remarkably straight trajectory every time.

The role of the brain in launching the tongue, therefore, is relatively simple. It's roughly analogous to the role of a medieval catapult operator: cranking the winch to increase the tension and sending the signal to let 'er fly. "Just because you have a complicated behavior doesn't necessarily mean you need a complicated control algorithm," says Nishikawa.

This theme was echoed in a talk by Michael Dickinson, who described recent work in his lab at the California Institute of Technology in Pasadena on the flight of fruit flies. He and his colleagues have been investigating the quick turns that the flies make in midflight to avoid collisions. The team's studies indicate that all that's needed to accomplish these dramatic turns—90 degrees in less than 50 milliseconds—is a modest tilt of the wing and a very slight change in the amplitude of each wing stroke. These adjustments are accomplished by steering muscles that alter the physical properties of the region of the thorax where the wing attaches (*Science*, 18 April 2003, p. 495). The brain can issue a few simple commands to the steering muscles, says Dickinson, "and what you get out is this beautiful maneuver."

Thinking outside the mouse

Asked to consider the future of their field, many neuroethologists see a mixed picture. On one hand, the field continues to attract young scientists and inspire new aproaches.

"We never have trouble getting students," says Kravitz. Several postdocs have joined his lab, which studies the roots of aggression, after doctoral studies on fruit fly genetics. Kravitz says the researchers were elated to find that their knowledge of genetics could be applied to study behavior.

This sort of influx also brings fresh approaches to bear on traditional problems. For more than 20 years, Kravitz has studied how hormones and neurotransmitters mediate aggressive behavior in lobsters. Thanks in part to the expertise of the newcomers, the lab is now doing more refined experiments in fruit flies, using genetic tools to tinker with specific signaling molecules in specific neurons.

Other researchers are also realizing the power of genetics—using DNA microarrays, for example, to hunt for genes that rev up when a fish moves up or down the dominance ladder. Still others are drawing on computational advances. Speakers at one symposium at the congress discussed how robots can be used to test models of the neural control of movement—as well as how lessons from neuroethology might be applied to design more lifelike robots.

At the same time, many neuroethologists say that they've come under increasing pressure to justify their work, as governmental grant agencies have focused more on applied research and have sought a return on their investment in major genome projects. The situation has pushed neuroscientists toward using a few select animal models, says John Hildebrand, who studies moth olfaction at the University of Arizona, Tucson.

The trend is reflected in an "overwhelming emphasis on mammals" at large neuroscience meetings such as the behemoth annual gathering of the Society for Neuroscience, Hildebrand says. A search through the program of this year's meeting, for example, turns up 2593 abstracts containing the word "rat" and 3554 containing "mouse" or "mice," but only 68 mentioning songbirds and 28 with electric fish.

Perhaps that's why many at the neuroethology congress say that the meeting has become a sort of refuge. "Every talk is like a celebration of a different model," says Hildebrand. Like many attendees, Dickinson says the meeting has become one of his favorites: "At Neuroethology you don't have to hide the fact that you're really interested in behavior, and that's kind of liberating in a way."

—GREG MILLER

CREDITS (TOP TO BOTTOM): A. KRISTOPHER LAPPIN/NORTHERN ARIZONA UNIVERSITY; JORG OESTREICH

Shocking gossip. Subtle changes in sodium channel genes have caused dramatic differences in communication signals among species of electric fish.

Cognitive Memory: Cellular and Network Machineries and Their Top-Down Control

Yasushi Miyashita

A brain-wide distributed network orchestrates cognitive memorizing and remembering of explicit memory (i.e., memory of facts and events). The network was initially identified in humans and is being systematically investigated in molecular/genetic, single-unit, lesion, and imaging studies in animals. The types of memory identified in humans are extended into animals as episodic-like (event) memory or semantic-like (fact) memory. The unique configurational association between environmental stimuli and behavioral context, which is likely the basis of episodic-like memory, depends on neural circuits in the medial temporal lobe, whereas memory traces representing repeated associations, which is likely the basis of semantic-like memory, are consolidated in the domain-specific regions in the temporal cortex. These regions are reactivated during remembering and contribute to the contents of a memory. Two types of retrieval signal reach the cortical representations. One runs from the frontal cortex for active (or effortful) retrieval (top-down signal), and the other spreads backward from the medial temporal lobe for automatic retrieval. By sending the top-down signal to the temporal cortex, frontal regions manipulate and organize to-be-remembered information, devise strategies for retrieval, and also monitor the outcome, with dissociated frontal regions making functionally separate contributions. The challenge is to understand the hierarchical interactions between these multiple cortical areas, not only with a correlational analysis but also with an interventional study demonstrating the causal necessity and the direction of the causality.

Since the pioneering observations on patient H.M., who developed a severe and selective deficit in the formation of explicit (or declarative) memory after a bilateral resection of the medial temporal lobe (i.e., the hippocampus and nearby regions), subsequent studies of patients have located the source of various types of impairment in explicit memory in many brain areas (*1*). Notably, although patients with localized frontal lobe lesions do not have an amnesia typically observed in patients with medial temporal lobe lesions, they do exhibit impairments in memory of temporal context or temporal order, memory of the source of facts or events, or metamemory (i.e., knowledge about one's memory capabilities and about strategies that can aid memory) (*2*–*4*).

The identified brain-wide distributed network, called here the cognitive memory system, is composed of three major subsystems, namely, the medial temporal lobe, the temporal cortex, and the frontal cortex (Fig. 1). Although the ultimate storage sites for explicit memories appear to be in the cortex [but see (*5*, *6*) for another strong position], the medial temporal lobe plays a critical enabling role necessary for storage to take place. Domain-specific cortical regions in the temporal lobes are reactivated during remembering and contribute to the contents of a memory. The reactivation process is mediated by various signals, such as the top-down signal from the prefrontal cortex or the backward signal from the limbic cortex. Frontal regions mediate the strategic attempts for retrieval and encoding and also monitor its outcome, with the dissociated frontal regions making functionally separate contributions. This large-scale cognitive network was initially identified in humans by using neuropsychology and functional imaging. However, molecular, cellular, and network components of this cognitive system have been systematically dissected by recent technical advancements, particularly in animal studies. These include cell type–restricted gene manipulations in mice, a combination of molecular biology and single-unit recording in monkeys, and a sophisticated scan design of event-related functional magnetic resonance imaging (fMRI) in humans. This review aims to integrate some recent discoveries on cellular and network machineries at multiple levels of complexity, which will help us to understand how this brain-wide network orchestrates cognitive memorizing and remembering.

Department of Physiology, University of Tokyo School of Medicine, Hongo, Tokyo 113-0033, Japan. E-mail: yasushi_miyashita@m.u-tokyo.ac.jp

Semantic-Like Memory and Episodic-Like Memory in Nonhuman Species

It is widely held that there are multiple types of memory that are mapped onto distinct anatomical circuits in the brain. Although various taxonomic frameworks for different types of memory have been proposed, most of them share a common generic form (Fig. 2) (*1*, *7*, *8*): A cardinal distinction lies between short-term (working) memory and long-term memory, and long-term memory is further divided into explicit (or declarative) and implicit (or nondeclarative) memory. Explicit memory is often further divided into semantic (fact) memory and episodic (event) memory: The former consists of facts about the world, such as the capital of the United States of America or that a horse has four legs and usually does not have wings (Fig. 1), whereas the latter was originally characterized as conscious recollection of specific events from one's personal past in humans (*7*). For investigations of cellular and neural-network mechanisms, some of these original definitions of memories provide difficult obstacles to studies in nonhuman animals. Regarding semantic cognition, there are lines of evidence that nonhuman animals do segment the world categorically into objects and that, although they lack verbal expression, they can display through their behavior that they ''know'' what these and other types of objects are (*9*, *10*). Direct physiological evidence is also available regarding neuronal correlates of object category representation in the cortex (*11*, *12*). Although previous approaches failed to capture some important semantic features of human memory, recent progress encourages us to investigate semantic-like memory, particularly in nonhuman primates (*9*, *13*). Also, some recent studies modeled episodic memory in animals as memory of ''when, where, and what'' event had occurred (*14*), or memory of an animal's own behavior (*15*). A detailed review of whether such episodic-like memory in animals meets the strict criteria required for human episodic memory (*10*, *16*, *17*) exceeds the scope of this article, and the following sections will focus on neurobiological bases of such episodic- or semantic-like memory.

Roles of Local Neuronal Circuits Within the Hippocampus

The medial temporal lobe, particularly the hippocampus, is a major site of multimodal convergence. It contains neurons that are sensitive to the configuration of many

environmental stimuli as well as to the behavioral context in which events occur (*5*, *18*–*20*). The hippocampus is considered to be critical to the formation of long-term explicit memories, which depends, most theories assume, on this type of configurational information and also on unique hippocampal neural circuits (*5*, *19*, *20*).

Anatomically, the hippocampus can be divided into several major subfields, namely, the dentate gyrus (DG), CA3, CA1, and subiculum, and each subfield has a distinct anatomical organization (*21*). Theoretical studies have proposed a complementary but computationally distinct mnemonic role for each of the hippocampal subfields. Some predictions were experimentally tested and confirmed. For example, the *N*-methyl-D-aspartate (NMDA) receptor, a type of glutamate receptor, has long been known to be critical to the formation of spatial memory in rats and mice (*22*), and some theories ascribed it to the NMDA receptors of the Schaffer collateral synapses on CA1 pyramidal cells (*20*, *23*). Cell type–restricted gene manipulations in mice, in which the NMDA receptor gene was ablated specifically in CA1 pyramidal cells, demonstrated that CA1 is indeed a major site involved in the storage of spatial reference memory (*24*, *25*).

However, some hypotheses on different aspects of the hippocampal circuit have not been tested. CA3 pyramidal cells are massively interconnected by recurrent collaterals (more than 10,000 synapses per pyramidal cell in the rat). This recurrent network in CA3 may be critical to various hippocampus-dependent memory functions, particularly "pattern completion" (*26*). The configuration of environmental stimuli and their behavioral context in daily life are unique and rarely repeated exactly, and thus the input patterns would be able to reactivate only a part of a stored memory and would be unable to activate the whole pattern unless a special computational process recovers the whole from a part. This "the-whole-from-a-part" process is called pattern completion, and the recurrent collateral network with modifiable synapses in CA3 was suggested to perform this computation. This conjecture was recently tested empirically in a genetically engineered mouse strain (called CA3-NR1 KO) where the NMDA receptor gene was ablated specifically in CA3 pyramidal cells of adult mice (*27*). Plasticity at CA3 recurrent collateral synapses was abolished in these mutants, whereas the plasticity at DG mossy fiber-CA3 synapses remained intact because the latter do not depend on the NMDA receptors. The mutant mice were normal in the acquisition and retrieval of spatial memory with repeated learning, tested with use of the Morris water maze. However, when memory of the hidden platform location was tested after removal of three of the four major extramaze cues, the mutants exhibited a deficit of memory retrieval. The lack of robustness to cue removal of the animal's spatial behavior was also mirrored in the responses of "place cells," which links NMDA CA3 receptors to behavior (*5*, *18*). Theoretically, pattern completion can contribute more to wide computational processes in episodic-like memory than shown above (*5*, *19*, *20*, *23*). One key feature, though not a sufficient criterion, of episodic-like memory is that it is acquired rapidly in a single trial and involves trial-specific information. A delayed matching-to-place (DMP) version of the Morris water maze task was used as a single-trial learning model, and an intrahippocampal infusion of the NMDA antagonist, D(–)-2 amino-5-phosphonovaleric acid (D-AP5), impaired the DMP task (*28*). This deficit in the DMP task can also be induced by NMDA receptor ablation restricted in CA3 pyramidal cells (*29*). In another example of an episodic-like memory test, rats encoded paired associates (flavors of food and their spatial locations) and recalled one item when cued by the other (*30*). When pairings of a particular food and its location were never repeated, ensuring unique "what-where" attributes that are other key features of episodic-like memory, intrahippocampal infusion of D-AP5 impaired memory encoding but not memory recall. Infusion of an AMPA receptor antagonist, CNQX, impaired both encoding and recall. In contrast, when paired associates were trained repeatedly over 8 weeks, the blockade of hippocampal AMPA receptors did not affect their recall. This indicates the differential roles of hippocampal and extrahippocampal neural circuits for nonrepeated and repeated learning.

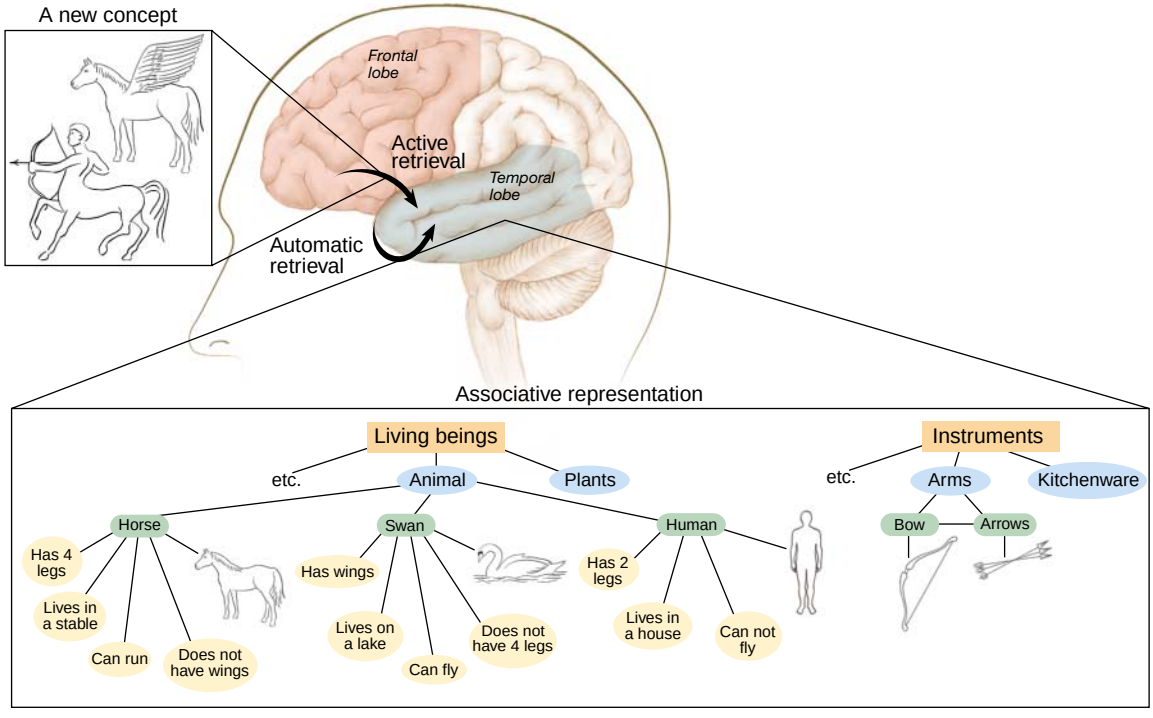


Fig. 1. A schematic drawing of the brain-wide network discussed in this article. Specific functional localizations are assigned on the basis of the observations on semantic-like memory in humans and monkeys (other types of memory are supported in different brain structures). The associative representations of long-term memory in the temporal cortex can be reactivated by either of two types of memory retrieval process; active retrieval process is supported by the signal that runs from the frontal cortex (top-down signal), whereas automatic retrieval signal is generated within the network of the temporal lobe and spreads backward. The cortical representations of experience or knowledge are schematically drawn as a semantic network (*51*). Neurobiologically they are likely mediated by pair-coding memory neurons that are created by a structural reorganization of neural circuits.

A similar conclusion was obtained by single-unit recording of the activity of hippocampal neurons in monkeys while they learned a new association between a scene and an eye movement direction in several successive trials (*31*). Hippocampal neurons changed their firing rate and stimulus selectivity during the location-scene learning, suggesting the involvement of these neurons in the initial formation of new associative memories.

These observations provide experimental supports for the theoretical considerations in the previous section. First, the unique configurational association between environmental stimuli and behavioral context, which is likely the basis of episodic-like memory, depends on distinct hippocampal neural circuits. Second, memory traces representing repeated associations are consolidated in neural circuits outside the hippocampus. With repetition, only the gist of the configurational information is reactivated and would be multiply represented among various contextual information that accompanied the events, eventually being associated with the subject's general knowledge as semantic-like memory (*8*, *13*). There is a debate on whether the hippocampus is necessary for retrieval of episodic memory per se (*5*–*8*), whereas it is agreed that other types of memory (e.g., semantic memory) are formed over time that do not depend on the hippocampal system for retrieval, though they may require it for acquisition. Third, as conjectured many times (*5*, *7*), human semantic and episodic memory may have evolved from the spatial knowledge and learning of the ancestors of humans. Studies on spatial memory have been providing many important works concerning the role of the hippocampus, which link molecular/genetic, single-unit, lesion, and functional imaging studies across various species. Many fine reviews of those studies are available and should be consulted (*5*, *32*, *33*).

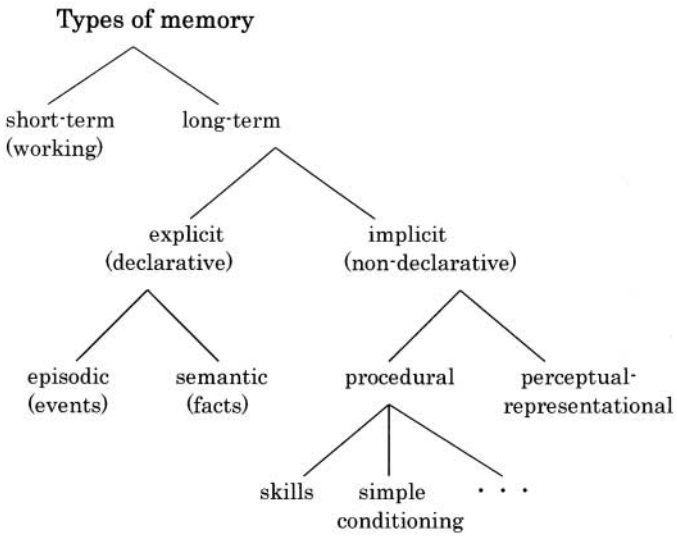


Fig. 2. Taxonomy of memory. Long-term memory is divided into explicit (declarative) memory and implicit (nondeclarative) memory. Implicit memory affects behavior without awareness. Explicit memory is further divided into semantic memory, representing general knowledge about the world, and episodic memory, representing personal knowledge of one's past. This generic form directly applies to the human memory system (*7*, *8*). Similar taxonomy would also apply to animal memory, though it may lack some characteristic features of human memory. Thus, the terms such as semantic-like or episodic-like memory are used in this article when referred to the animal memory systems.

Cortical Representation of Associative Memory: Temporal Association Mechanism

Where, outside the hippocampus, are memory traces representing such repeated associations located, and how are they organized? Several lines of evidence suggest that they are organized in the higher order association cortex for semantic-like memory (*13*, *34*–*36*) as well as in subcortical structures, such as the basal ganglia, for some nondeclarative memory (*35*, *37*). The neuronal correlates of associative long-term memory were first reported in the monkey inferior temporal (IT) cortex (*38*, *39*). In order to investigate the neurobiological basis of semantic-like memory, monkeys were trained to perform the pair-association memory task, a well-known neuropsychological test that is widely used for assessment of dysfunction of the medial temporal lobe system in humans. Single neurons in the IT cortex were found to create linkages between representations of operationally associated but physically unrelated visual stimuli. These associations turned out to be formed by temporally correlated activity in the network, as also confirmed by later studies (*40*, *41*). This work opened the door to neurobiological investigations of the cortical semantic-like memory network by reducing a complex network into its elementary associative links between two objects and then by seeking molecular and morphological machineries underlying such elementary associative links (*13*, *36*).

Physiological mapping and comparison of memory representations in different temporal cortical regions [for anatomical definitions, see Fig. 3A and (*42*, *43*)] revealed that the percentage of memory neurons (the ''pair-coding'' neurons) that encode both paired associates was significantly higher in the limbic cortex (area 36, 33%) than in the adjoining neocortex (area TE, 4.9%) (*44*). The functional architecture was also different between TE and area 36: Pair-coding neurons were found to form a local cluster of about 1 mm in area 36 but were distributed more sparsely in TE. Thus, although neurons in both areas acquire stimulus selectivity through learning, the association between representations of visual paired associates proceeds forward within the IT cortex in multiple steps.

Molecular and morphological bases of neural circuit reorganization. It has long been hypothesized that the memory engrams of declarative knowledge in the cortex develop with the structural reorganization of neural circuits (*34*–*36*). This reorganization of neural circuits would be accomplished through a cellular program of gene expression leading to increased protein synthesis and then to an alteration of synaptic connections (*45*). This hypothetical framework has been primarily investigated in invertebrates and lower mammals.

Recently, this hypothesis has been tested with the pair-association task for semantic-like memory in a series of molecular biological studies carried out in monkeys. Up-regulation of mRNAs encoding proteins, thought to be involved in structural reorganization, occurs during the formation of the pair-association memory in a narrow cortical area where pair-coding neurons are physiologically located (Fig. 3, B to D) (*46*). In these studies, reverse transcription polymerase chain reaction mRNA quantitation was combined with an experimental strategy using split-brain monkeys, in which the anterior commissure and the entire extent of the corpus callosum were transected. This preparation enables the comparison of mRNA expressions in both hemispheres (the pair-association hemisphere and the control hemisphere) within the same monkey, thereby eliminating genetic and cognitive variations between individuals. mRNA encoding the gene of brain-derived neurotrophic factor (BDNF) and an immediate-early gene, *zif268*, were found to increase in area 36. The spatial distribution of mRNAs was also visualized with the use of in situ hybridization. BDNF mRNA–positive cells and ZIF mRNA–positive cells accumulated as a patchy cluster in area 36, extending for at least 0.4 mm along the anterior-posterior axis.

BDNF is considered to mediate activity-dependent synaptic plasticity, even in mature nervous systems, and the BDNF Val^{66}met

polymorphism was demonstrated to affect activity-dependent secretion of BDNF and human episodic memory (*47*). Because *zif 268* encodes a transcription factor (a protein that binds to DNA and controls a transcription of other genes), its expression may trigger a cascade of gene activation that leads to cellular events underlying the neuronal circuit reorganization. This hypothesis was tested by a morphological approach (*48*). In monkeys trained in a pair-association task, neurons selective to learned pictures formed a focal patch in area 36 (''hot spot''). Three types of retrograde tracer were injected into area 36. One tracer was injected into the hot spot. Two different tracers were injected into two regions adjacent but outside the hot spot. Then, the distribution of retrogradely labeled neurons and electrophysiologically recorded neurons was compared in TE. Picture-selective neurons in TE projected less divergently and more specifically to the hot spot than other neurons, suggesting that, after visual learning, axonal arbors originating from learning-related neurons are pruned to connect specifically to the patch in area 36 whereas those from other neurons retain their divergence. An interesting conjecture is that this learning-induced axonal pruning outside the hot spot is related to the enhanced local extension of axonal arbors within the hot spot (*48*), leading to the reorganization of local networks in the hot spot that is detected electrophysiologically as a change in neuronal stimulus selectivity, that is, the emergence of pair-coding neurons.

Activation of Memory Representations: Active Versus Automatic Retrieval

The associative long-term memory stored in the temporal cortex can be retrieved by either of two types of memory retrieval process: one occurs when we need no effort to recall and the other when we have to strive toward a successful recall. We refer to the former as automatic retrieval and to the latter as active, or effortful, retrieval (*13*). The concept of active retrieval stems from the controlled processing in cognitive theories (*49*, *50*), which was characterized to be capacity-limited and operating when the task cannot be accomplished through automated stimulus-response mapping. I now examine evidence supporting the hypothesis that automatic retrieval and active retrieval are supported by retrieval signals generated within the temporal lobe network and by signals that run from the frontal cortex to the temporal cortex, respectively (*13*).

Automatic retrieval signal: backward spread of memory signal in the temporal lobe. The cognitive theory of semantic network postulates retrieval of an item as an activation of a corresponding node in the network (*51*). The neural correlate of such a node activation was first reported in a pair-association task (*39*). The response was referred to as a pair-recall response. Then, by using a modified pair-association task (PA with a color switch task), Naya *et al.* showed that this pair-recall response indeed corresponds to the recall of the target in the subject's mind, because IT neurons started firing immediately after a color switch that signaled the necessity and timing of memory retrieval during a delay period (*52*). IT neurons also stopped firing immediately after another color switch that signaled the retrieval of other memorized items. Recently, propagation of the pair-recall activity in the temporal lobe has been investigated (*53*). The onset of the pair-recall activity was much earlier and the activity developed more rapidly in area 36 neurons than in TE neurons. The median retrieval time was over 300 ms longer in TE than in area 36. Therefore, memory retrieval signals appeared first in the limbic cortex (area 36), after which neocortical (TE) neurons were gradually recruited to represent the sought target. Thus, the mnemonic information that was extracted from long-term storage spreads backward from the limbic cortex to the neocortex in the temporal lobe.

Top-down signaling appears when active retrieval is required. A clinical case study highlights active retrieval in humans and provides a clue to an experimental model with which active retrieval can be investigated (*54*). An epileptic patient who had undergone posterior callosotomy (i.e., partial disconnection of the commissural fibers connecting the left and right cerebral hemispheres) was presented a word in his left visual field. He could not read the name of

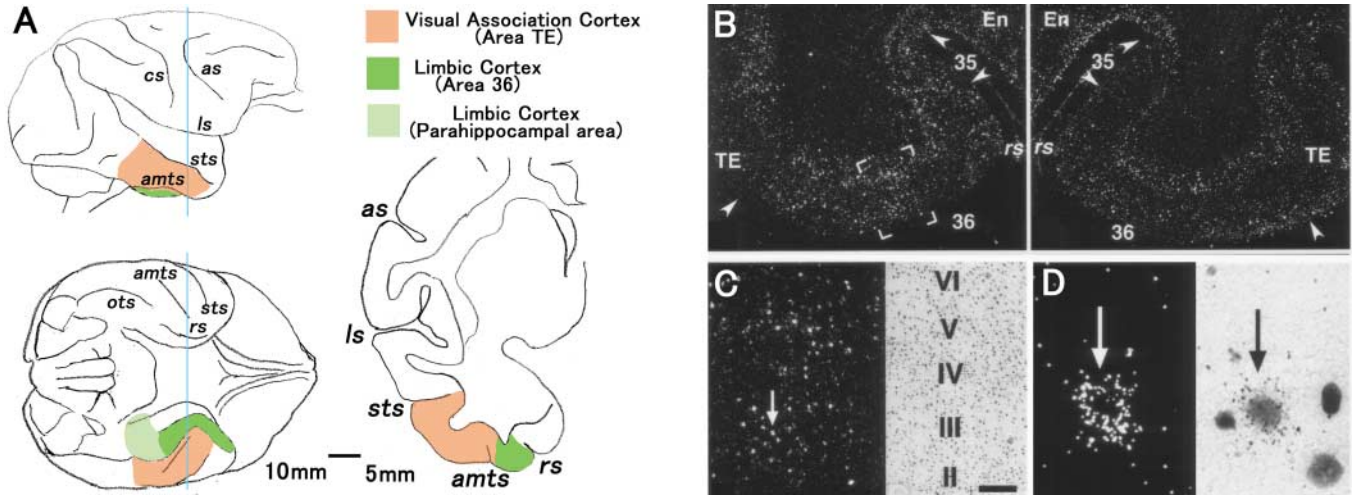


Fig. 3. Neocortical and limbic cortical areas in the temporal lobe of monkeys, and local BDNF induction in memory formation. (**A**) Anatomical relationship between neocortical area (area TE, orange) and limbic areas (area 36, green; parahippocampal area, light green). Top, lateral view of a monkey brain. Bottom left, bottom view. Bottom right, coronal section cut at the line shown in the bottom view. rs, rhinal sulcus; amts, anterior middle temporal sulcus; sts, superior temporal sulcus; ots, occipitotemporal sulcus; cs, central sulcus; ls, lateral sulcus; as, arcuate sulcus. Scale-bar: 10 mm, bottom left; 5 mm, bottom right (B to D) BDNF induction in monkey temporal cortex (area 36) during the formation of pair-association (PA) memory. (**B**) In situ hybridization of BDNF mRNA in the inferior temporal gyrus of the split-brain monkey, the PA hemisphere (left), and the control hemisphere (right). BDNF mRNA accumulated in a patch in area 36 of the PA hemisphere (framed area), but not of the control hemisphere. (**C**) The framed area in (B) is enlarged. (Left) BDNF mRNA–positive cells were observed in layers V/VI and in layers II/III. (Right) Corresponding area in an adjacent Nissl section. (**D**) BDNF mRNA–positive cells. Cell marked by arrow in (C) is enlarged and shown in dark field (left) and bright field (right). En, entorhinal cortex; 35, area 35; 36, area 36; TE, area TE; rs, rhinal sulcus. Arrowheads mark the boundaries between different cortical areas. [Modified from (*46*)]

it, consistent with the fact that the bottom-up visual information could not directly reach the language areas in the left hemisphere. However, he claimed to ''see'' its image in his mind. He was eventually able to answer the name by using inferential strategies based on his mental image. His limited ability suggests that his right hemisphere was transmitting to his left hemisphere, through the commissural fibers of the prefrontal cortex, semantic information about the stimulus but not the actual stimulus. This posterior-split-brain paradigm was combined with the associative memory task in monkeys (*55*). In a posterior-split-brain monkey, in which only the anterior corpus callosum remains intact and other commissural fibers are surgically transected, the cortex receives bottom-up visual information only from the contralateral visual field. In this preparation, long-term memory acquired through pair-association learning does not transfer interhemispherically via the anterior corpus callosum; nonetheless, when the visual cue is presented to one hemisphere, the anterior callosum can instruct the other hemisphere to retrieve the correct stimulus specified by the cue. Thus, although visual long-term memory is stored in the temporal cortex, memory retrieval is under the executive control of the frontal cortex.

A direct proof of the existence of top-down signaling was provided by single-unit recordings from the temporal cortex of posterior-split-brain monkeys (*56*). A considerable number of IT neurons did indeed receive top-down signals from the frontal cortex as well as bottom-up signals from the retina. The response latency was longer in the top-down input, reflecting the multisynaptic conduction delay within the frontal cortex. The top-down signals conveyed a categorical feature of the stimulus rather than a physical feature of it, consistent with a report on prefrontal neuronal responses in a stimulus categorization task (*57*) and in working memory tasks (*3*). In summary, the partial split-brain studies in humans and monkeys revealed the events occurring during the active retrieval process, in which top-down signals from the frontal cortex trigger the activation of memory representations in the temporal cortex.

Imaging studies in humans further confirmed that the frontal cortex plays a key role in the active retrieval process. Activation of the frontal cortex during memory retrieval is widely observed in functional neuroimaging studies using various psychological paradigms and test modalities, including recognition tests, word-stem tasks, word-fragment tasks, paired associates tasks, free recall, and recency judgment [for review, see (*58*–*61*)]. Although some initial neuroimaging studies in humans failed to detect the activation of the posterior domain-specific cortical regions as the target of recapitulation mechanisms, recent event-related fMRI studies detected recapitulation effects in these content-holding posterior cortices (*33*, *62*–*65*). Furthermore, within the ventral visual processing stream, content-based modulation was specific to late object-responsive regions, sparing the areas that process more primitive retinotopically mapped visual features (*64*).

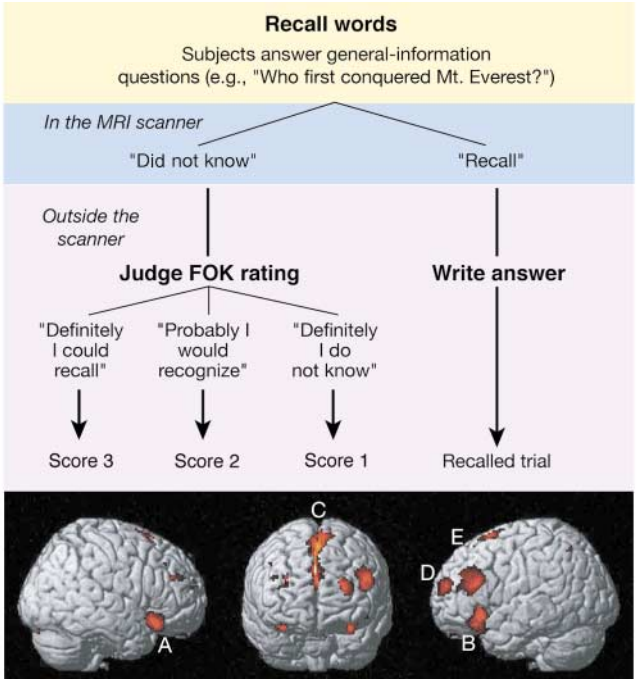


Fig. 4. Controlled memory retrieval system in the prefrontal cortex, as exemplified in fMRI activation in a FOK task. (**Top**) Experimental procedures of a FOK task. FOK is a subjective sense of knowing an item or a word before recalling it, and is a well-established tool for assessing the metamemory system. In this experiment, subjects were required to recall word answers to general-information questions, such as, "Who established the company SONY?" or "Who first conquered Mount Everest?" during fMRI scans. They indicated whether they recalled the target words or not. Then they judged their degree of FOK to the nonrecalled questions on a multiple-point scale: that is, in a three-point scale: 3, "I definitely could recall the answer if given hints or more time"; 2, "I probably would recognize the answer"; and 1, "I definitely did not know." Each trial was sorted into trial type (Recalled, Score3, Score2, and Score1) according to the participant's judgment and was subjected to event-related fMRI analysis. (**Bottom**) Regions showing stronger fMRI activity when the subjects had a greater FOK. A and B, inferior frontal gyrus (BA 47); C, anterior cingulate cortex/supplementary motor area (BA 32/24/6); D and E, MFG (BA 10, BA 46/9). [Modified from (*68*)]

Cognitive Control and Top-Down Signaling

What types of cognitive process in the frontal lobes are sending the top-down signals? Patients with localized frontal lobe lesions are impaired in such tasks that tap into, for example, memory of temporal order (recency memory), source memory, or metamemory (*1*–*4*, *66*, *67*). There are many functional imaging literatures on prefrontal contributions to controlled memory retrieval, but relatively small number of reports directly investigated the above types of tasks [for reviews, see (*58*–*61*)]. A flavor of recent attempts to break down such complex cognitive functions into elementary processes can be seen in an example of a feeling-of-knowing (FOK) task (*68*, *69*). FOK is a subjective sense of knowing an item or a word before recalling it and is a well-established tool for assessing the metamemory system (Fig. 4, top) (*70*). Event-related fMRI revealed multiple frontal regions that showed stronger activity when the subjects had a greater FOK, including the bilateral inferior frontal gyri (BA 47), left middle frontal gyrus (MFG) (BA 46/9), frontopolar area (BA 10), and anterior cingulate/supplementary motor areas (BA 32/24/6) (Fig. 4, bottom). In human neuroimaging, the identification and dissociation of distinct frontal regions have been extensively pushed forward by using simpler, controlled memory-retrieval tasks, such as an episodic recognition task or a source retrieval task (*58*–*61*). Cognitive and neuropsychological theories suggest several controlled processes common for these controlled retrieval tasks and the FOK task (*49*, *50*, *70*, *71*). The retrieval-cue-specification process systematically analyzes possible semantic relationships between the retrieval cue and the known characteristics of the potential targets. If self-generated cues trigger necessary semantic knowledge or explicit source recollection unique to the targets, then an appropriate memory judgment can be made. The recollection monitoring process evaluates the products of memory retrieval with respect to their relevance to the retrieval demands. In verbal or verbalizable tasks, phonological maintenance and rehearsal processes are recruited. However, the specific functional roles of the identified

frontal regions related to controlled memory retrieval are currently still under debate (*58*–*61*, *72*), though many researchers suggest that the anterior portion of the left MFG near BA10 is related to high-level retrieval strategy (*61*). At present, a similar fMRI study of these complex cognitive tasks in monkeys looks to be simply a dream. But I believe that a powerful approach to test the suggested functional roles of the frontal regions will be provided in monkey experiments by the combination of monkey fMRI (*73*, *74*) and reversible cortical inactivation with a local drug injection.

Conclusions

Neuroimaging studies in humans identified a large brain-wide network of cognitive memory. A subsystem in the parietal cortex (*58*–*60*, *75*) was not discussed in this article in spite of its importance, because there are few animal data for possible cellular and network machineries that substantiate suggested parietal functions. Initial neuroimaging studies on memory, particularly positron emission tomography and fMRI studies, often emphasized only the activation of the frontal regions without detectable activation in the posterior cortices, which apparently disagreed with neuropsychological literature [for historical review, see (*1*)]. However, a consensus appears to be emerging. Recent event-related fMRI successfully detected activation in the posterior cortices in the controlled retrieval tasks (*62*–*65*) and spatial navigation tasks (*33*). The frontal regions are critically involved in manipulating and organizing to-be-remembered information and in devising strategies and monitoring for retrieval, although the frontal regions themselves may not be intimately involved in the binding of information into long-term memory (*1*–*4*, *58*–*61*). This consensus was promoted by cellular- and network-level observations in mice, rats, and monkeys. The neuronal representations of mnemonic contents in the medial temporal lobe and the posterior cortices, as well as the interacting signals between and within these structures, have been analyzed by genetic and single-unit recording approaches.

The existence of top-down signaling from the frontal cortex to the temporal cortex was directly demonstrated in monkeys (*55*, *56*), and neuronal activities were observed in the monkey frontal cortex in relation to various cognitive memory tasks (*3*, *57*, *76*, *77*). However, we still lack basic knowledge on how these frontal neurons are incorporated into network machineries that send out the top-down signal and/or support the cognitive processes identified by neuropsychology and neuroimaging in humans (*1*–*4*, *76*). The challenge is to clarify hierarchical interactions or couplings between multiple cortical areas as initially demonstrated by an effective connectivity analysis or a correlation analysis with fMRI in humans (*78*, *79*). Obviously, the time resolution of these hemodynamic analyses is low when compared with cell-level signaling. Moreover, because these analyses are correlational, only an interventional study will settle the causal relation and the direction of the causality. Thus, the necessary technical breakthrough would be, I believe, to construct an animal model of source memory or metamemory and to apply an interventional approach such as a reversible inactivation with a genetic engineering or a local drug injection. When our knowledge on these cortical interactions advances, we may be able to understand the neurobiological basis of our metacognition.

References and Notes

1. L. R. Squire, D. L. Schacter, *Neuropsychology of Memory* (Guilford, New York, ed. 3, 2002).
2. D. T. Stuss, D. F. Benson, *The Frontal Lobes* (Raven, New York, 1986).
3. J. M. Fuster, *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe* (Lippincott-Raven, Philadelphia, 1997).
4. M. Petrides, in *Handbook of Neuropsychology*, F. Boller, J. Grafman, Eds. (Elsevier, Amsterdam, 2000).
5. J. O'Keefe, L. Nadel, *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, Oxford, 1978).
6. L. Nadel, M. Moscovitch, *Curr. Opin. Neurobiol.* **7**, 217 (1997).
7. E. Tulving, *Elements of Episodic Memory* (Oxford Univ. Press, New York, 1983).
8. L. R. Squire, C. E. Stark, R. E. Clark, *Annu. Rev. Neurosci.* **27**, 279 (2004).
9. W. A. Roberts, *Principles of Animal Cognition* (McGraw Hill, Boston, 1998).
10. N. S. Clayton, T. J. Bussey, A. Dickinson, *Nat. Rev. Neurosci.* **4**, 685 (2003).
11. E. K. Miller, D. J. Freedman, J. D. Wallis, *Philos. Trans. R. Soc. London Ser. B* **357**, 1123 (2002).
12. N. Sigala, N. K. Logothetis, *Nature* **415**, 318 (2002).
13. Y. Miyashita, T. Hayashi, *Curr. Opin. Neurobiol.* **10**, 187 (2000).
14. N. S. Clayton, A. Dickinson, *Nature* **395**, 272 (1998).
15. R. R. Hampton, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 5359 (2001).
16. R. G. M. Morris, *Philos. Trans. R. Soc. London Ser. B* **356**, 1453 (2001).
17. P. R. Hampton, B. L. Schwartz, *Curr. Opin. Neurobiol.* **14**, 192 (2004).
18. J. O'Keefe, J. Dostrovsky, *Brain Res.* **34**, 171 (1971).
19. H. Eichenbaum, *Nature Rev. Neurosci.* **1**, 41 (2000).
20. E. T. Rolls, *Hippocampus* **6**, 601 (1996).
21. D. G. Amaral, M. P. Witter, *Neuroscience* **31**, 571 (1989).
22. T. V. Bliss, G. L. Collingridge, *Nature* **361**, 31 (1993).
23. B. L. McNaughton, R. G. M. Morris, *Trends Neurosci.* **10**, 408 (1987).
24. J. Z. Tsien *et al.*, *Cell* **87**, 1317 (1996).
25. M. Mayford *et al.*, *Science* **274**, 1678 (1996).
26. D. Marr, *Philos. Trans. R. Soc. London Ser. B* **262**, 23 (1971).
27. K. Nakazawa *et al.*, *Science* **297**, 211 (2002); published online 30 May 2002 (10.1126/science.1071795).
28. R. J. Steele, R. G. M. Morris, *Hippocampus* **9**, 118 (1999).
29. K. Nakazawa *et al.*, *Neuron* **38**, 305 (2003).
30. M. Day, R. Langston, R. G. Morris, *Nature* **424**, 205 (2003).
31. S. Wirth *et al.*, *Science* **300**, 1578 (2003).
32. M. W. Brown, J. P. Aggleton, *Nature Rev. Neurosci.* **2**, 51 (2001).
33. N. Burgess, E. Maguire, J. O'Keefe, *Neuron* **35**, 625 (2002).
34. L. R. Squire, S. Zola-Morgan, *Science* **253**, 1380 (1991).
35. M. Mishkin, W. A. Suzuki, D. G. Gadian, T. G. Aigner, *Philos. Trans. R. Soc. London Ser. B* **352**, 1461 (1997).
36. Y. Miyashita, *Annu. Rev. Neurosci.* **16**, 245 (1993).
37. R. A. Poldrack *et al.*, *Nature* **414**, 546 (2001).
38. Y. Miyashita, *Nature* **335**, 817 (1988).
39. K. Sakai, Y. Miyashita, *Nature* **354**, 152 (1991).
40. V. Yakovlev, S. Fusi, E. Berman, E. Zohary, *Nature Rev. Neurosci.* **1**, 310 (1998) .
41. C. A. Erickson, R. Desimone, *J. Neurosci.* **19**, 10404 (1999).
42. W. A. Suzuki, D. G. Amaral, *J. Comp. Neurol.* **463**, 67 (2003).
43. K. S. Sallem, K. Tanaka, *J. Neurosci.* **16**, 4757 (1996).
44. Y. Naya, M. Yoshida, Y. Miyashita, *J. Neurosci.* **23**, 2861 (2003).
45. C. H. Bailey, E. R. Kandel, *Annu. Rev. Physiol.* **55**, 397 (1993).
46. W. Tokuyama, H. Okuno, T. Hashimoto, Y. X. Li, Y. Miyashita, *Nature Rev. Neurosci.* **3**, 1134 (2000).
47. M. F. Egan *et al.*, *Cell* **112**, 257 (2003).
48. M. Yoshida, Y. Naya, M. Miyashita, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4257 (2003).
49. R. M. Shiffrin, W. Schneider, *Psychol. Rev.* **91**, 269 (1984).
50. P. W. Burgess, T. Shallice, *Memory* **4**, 359 (1996).
51. J. R. Anderson, *Cognitive Psychology and Its Implications* (Freeman, New York, ed. 4, 1995).
52. Y. Naya, K. Sakai, Y. Miyashita, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 2664 (1996).
53. Y. Naya, M. Yoshida, Y. Miyashita, *Science* **291**, 661 (2001).
54. J. J. Sidtis, B. T. Volpe, J. D. Holtzman, D. H. Wilson, M. S. Gazzaniga, *Science* **212**, 344 (1981).
55. I. Hasegawa, T. Fukushima, T. Ihara, Y. Miyashita, *Science* **281**, 814 (1998).
56. H. Tomita, M. Ohbayashi, K. Nakahara, I. Hasegawa, Y. Miyashita, *Nature* **401**, 699 (1999).
57. D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, *Science* **291**, 312 (2001).
58. M. D'Esposito, *Semin. Neurol.* **20**, 487 (2000).
59. R. L. Buckner, M. E. Wheeler, *Nature Rev. Neurosci.* **2**, 624 (2001).
60. A. D. Wagner, L. Davachi, *Curr. Biol.* **11**, R964 (2001).
61. P. C. Fletcher, R. N. Henson, *Brain* **124**, 849 (2001).
62. A. Ishai, L. G. Ungerleider, J. V. Haxby, *Neuron* **28**, 979 (2000).
63. K. M. O'Craven, N. Kanwisher, *J. Cognit. Neurosci.* **12**, 1013 (2000).
64. M. E. Wheeler, S. E. Petersen, R. L. Buckner, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11125 (2000).
65. I. Kahn, L. Davachi, A. D. Wagner, *J. Neurosci.* **24**, 4172 (2004).
66. J. S. Janowsky, A. P. Shimamura, L. R. Squire, *Neuropsychologia* **27**, 1043 (1989).
67. B. Milner, P. Corsi, G. Leonard, *Neuropsychologia* **29**, 601 (1991).
68. H. Kikyo, K. Ohki, Y. Miyashita, *Neuron* **36**, 177 (2002).
69. A. Maril, J. S. Simons, J. P. Mitchell, B. L. Schwartz, D. L. Schacter, *Neuroimage* **18**, 827 (2003).
70. A. Koriat, R. Levy-Sadot, *J. Exp. Psychol. Learn. Mem. Cognit.* **27**, 34 (2001).
71. D. L. Schacter, K. A. Norman, W. Koutstaal, *Annu. Rev. Psychol.* **49**, 289 (1998).
72. I. G. Dobbins, H. Foley, D. L. Schacter, A. D. Wagner, *Neuron* **35**, 989 (2002).
73. N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, A. Oeltermann, *Nature* **412**, 128 (2001).
74. K. Nakahara, T. Hayashi, S. Konishi, Y. Miyashita, *Science* **295**, 1532 (2002).
75. M. D. Rugg, A. P. Yonelinas, *Trends Cognit. Sci.* **7**, 313 (2003).
76. E. K. Miller, J. D. Cohen, *Annu. Rev. Neurosci.* **24**, 167 (2001).
77. M. Ohbayashi, K. Ohki, Y. Miyashita, *Science* **301**, 233 (2003).
78. E. Koechlin, C. Ody, F. Kouneiher, *Science* **302**, 1181 (2003).
79. K. Sakai, P. E. Passingham, *Nature Rev. Neurosci.* **6**, 75 (2003).
80. This work was supported by a Grant-in-Aid for Specially Promoted Research (14002005) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Language and the Origin of Numerical Concepts

Rochel Gelman* and C. R. Gallistel*

Reports of research with the Pirahã and Mundurukú Amazonian Indians of Brazil lend themselves to discussions of the role of language in the origin of numerical concepts. The research findings indicate that, whether or not humans have an extensive counting list, they share with nonverbal animals a language-independent representation of number, with limited, scale-invariant precision. What causal role, then, does knowledge of the language of counting serve? We consider the strong Whorfian proposal, that of linguistic determinism; the weak Whorfian hypothesis, that language influences how we think; and that the "language of thought" maps to spoken language or symbol systems.

Intuitively, our thoughts are inseparable from the words in which we express them. This intuition underlies the strong form of the Whorfian hypothesis, namely, that language determines thought (aka "linguistic determinism"). Many cognitive scientists find the strong hypothesis unintelligible and/or indefensible (*1*), but weaker versions of it, in which language influences how we think, have many contemporary proponents (*2*, *3*).

The strong version rules out the possibility of thought in animals and humans who lack language, although there is an abundant experimental literature demonstrating quantitative inference about space, time, and number in preverbal humans (*4*), in individuals with language impairments (*5*), and in rats, pigeons, and insects (*6*). Another problem is the lack of specific suggestions as to how exposure to language could generate the necessary representational apparatus. It would be wonderful if computers could be made to understand the world the way we do just by talking to them, but no one has been able to program them to do this. This failure highlights what is missing from the strong form of the hypothesis, namely, suggestions as to how words could make concepts take form out of nothing.

The antithesis of the strong Whorfian hypothesis is that thought is mediated by language-independent symbolic systems, often called the language(s) of thought (*7*). Under this account, when humans learn a language, they learn to express in it concepts already present in their prelinguistic system(s). Their prelinguistic systems for representing the world are language-like only in that they are compositional: Larger, more complex meanings (concepts) are built up by the combination of elementary meanings.

Psychology and Cognitive Science, Rutgers University, 152 Frelinghuysen Road, Piscataway, NJ 08854. USA.

*To whom correspondence should be addressed. E-mail: rgelman@ruccs.rutgers.edu (R.G.); Galliste@ruccs.rutgers.edu (C.R.G.)

Recently reported experimental studies (*8*, *9*) with innumerate Pirahã and Mundurukú Indian subjects from the Brazilian Amazonia give evidence regarding the role of language in the development of numerical reasoning. Either the subjects in these reports have no true number words (*8*, *10*) or they have consistent, unambiguous words for one and two and more loosely used words for three and four (*9*). Moreover, they do not overtly count, either with number words or by means of tallies. Yet, when tested on a variety of numerical tasks—naming the number of items in a stimulus set, constructing sets of equivalent number, judging which of two sets is more numerous, and mental addition and subtraction—these subjects gave results indicative of an imprecise nonverbal representation of number, with a constant level of imprecision, measured by the Weber fraction. The Weber fraction for these subjects is roughly comparable to that of numerate subjects when they do not rely on verbal counting. In one of the reports, the stimulus sets had as many as 80 items, so the approximate representation of number in these subjects extends to large numbers.

Among the most important results in these reports are those showing simple arithmetic reasoning—mental addition, subtraction, and ordering. These findings strengthen the evidence that humans share with nonverbal animals a language-independent representation of number, with limited, scale-invariant precision, which supports simple arithmetic computation and which plays an important role in elementary human numerical reasoning, whether verbalized or not (*5*, *11*–*13*). Contrary to (*8*) and to reports in the secondary media, the results do not support the strong Whorfian view that a concept of number is dependent on natural language for its development. Indeed, they are evidence against it. The results are, however, consistent with the hypothesis that learning to represent numbers by some communicable notation (number words, tally marks, numerals) might facilitate the routine recognition of exact numerical equality.

These reports suggest that people with extremely limited or no verbal counting have the same nonverbal representation of number as do subjects with a fluent, well-developed verbal counting system. The long-established and robust symbolic size and distance effect is a principal line of evidence for this representation and for its importance in discussions of verbal numerical reasoning: Numerate subjects judge the ordering of symbolized number with ease, but they have no insight into how they do so. Most are surprised to learn that it takes them longer to decide that $3 > 2$ than it does to decide that $5 > 2$, whether the questions are posed symbolically ($3 >? 5$) or with arrays instantiating the numbers (Fig. 1, A and B). The reaction time for judgments of numerical order is a function of the ratio between the numbers being judged. The function is the same in monkeys as in numerate adults (Fig. 1, C and D).

The symbolic size and distance effects are generally taken to indicate that the determination of numerical order by the brain depends on imprecise mental magnitudes. These are hypothesized variables in the brain that vary systematically with number (and other quantitative dimensions of experience) and that form the basis for the subjective sense of magnitude. They are called mental or subjective magnitudes to distinguish them from the objective magnitudes that they represent. The mental magnitudes for repeated instantiations of an objective magnitude vary, forming what communications engineers call a signal distribution. The wider these distributions, the more imprecise are the representations. The extent to which two signal distributions overlap determines the likelihood of confusion about which distribution a signal belongs to, that is, which objective magnitude generated it. It is assumed that the extent of overlap between two signal distributions is determined by the ratio between the corresponding objective magnitudes (Fig. 2). The greater the likelihood of confusion, the more processing time is required to determine the proper distribution. This is the generally accepted explanation for the symbolic size and distance effects. It ties basic arithmetic reasoning (order judgments) with numerical symbols to an imprecise nonverbal representation of number.

On the non-Whorfian account, the mental magnitudes that represent number are an example of elementary nonlinguistic representations (meanings) for which numerate

subjects have learned words. Subjects believe that the property denoted by "three" may be added to the property denoted by "two" to obtain the property denoted by "five" because this is already true for the prelinguistic concepts to which the words refer and from which they derive their meaning. That is, the mental magnitude that represents three may be mentally added to the mental magnitude that represents two to get the mental magnitude that represents five. Plausibly, the language learner comes to believe that those words have those meanings, precisely because she observes that their use is consistent with those meanings. Children hear, "Three and two are five" but not "Cow and big are red." From the syntactic frames in which words occur, much may be inferred about their referents (*14*, *15*).

In showing that subjects with no verbal counting system have a concept of approximate numerical magnitude like that of numerate subjects, these reports support the non-Whorfian view for the origins of our concept of number. However, there is more to the story. Numerate subjects have a strong intuition of exact numerical equality. Two plus two is exactly four, not roughly four, and the square root of two is not exactly equal to the proportion between any two count numbers, that is, to any rational number, although a rational number that is as close as one wishes may readily be found. This aspect of the meaning of number words is not readily explained by the assumption that it is the reference to imprecise mental magnitudes that gives number words their meanings. When the non-numerate subjects in these reports matched a set of four items to a set of five, or judged that 6 – 3 = 2, they gave evidence of being indifferent to exact numerical equality, an indifference not seen in numerate control subjects. Thus, the reports suggest that the learning of number words either creates a concept of exact numerical equality (a strong Whorfian hypothesis), or mediates the expansion of such a concept (a weaker Whorfian hypothesis), or directs attention to such a concept (a non-Whorfian hypothesis).

A current hypothesis of the second (weak Whorfian) kind is the two-systems hypothesis, which is that, in addition to the approximate representation of numerical magnitude, there is a second prelinguistic representation, limited to numbers from one to four (*16*). Because this second, small-number-only system is discrete and precise, exact equality is intrinsic to the representation. It comes from the identity of the representing symbols (e.g., II ≡ II). On these accounts, the acquisition of a verbal counting system mediates the extension of the notion of exact equality to our concept of numbers greater than four.

The reports from experiments with nonnumerate subjects do not offer much support for two-system hypotheses, because the innumerate subjects represent three and four imprecisely—and, arguably, even one and two. This finding is also problematic for closely related and long-popular hypotheses that postulate perceptlike representations of the numbers one to four (*17*, *18*). If these subjects have precise or percept-like prelinguistic representations of three and four, then, curiously, they have no words that refer unambiguously to them.

There are at least two conceptual problems with dual-representation hypotheses. First, if the words for the numbers one to four derive their meaning from discrete, noise-free prelinguistic symbols or percepts, then why are the symbolic size and distance effects seen in this range? (Recall that these effects are assumed to derive from the imprecision of the mental magnitude representation of number.) Second, the compositionality of number concepts is a sine qua non. If the brain represents three and seven in fundamentally different ways, how can it compose them arithmetically (order them, add them, etc.)? What representational form do the resulting hybrids have? This is particularly puzzling when two numbers beyond the discrete and precise range are subtracted to yield a number inside it, as in 7 – 5 = 2.

Nonetheless, reports of subjects who appear indifferent to exact numerical equality even for small numbers, and who also do not count verbally, add weight to the idea that learning a communicable number notation with exact nu-

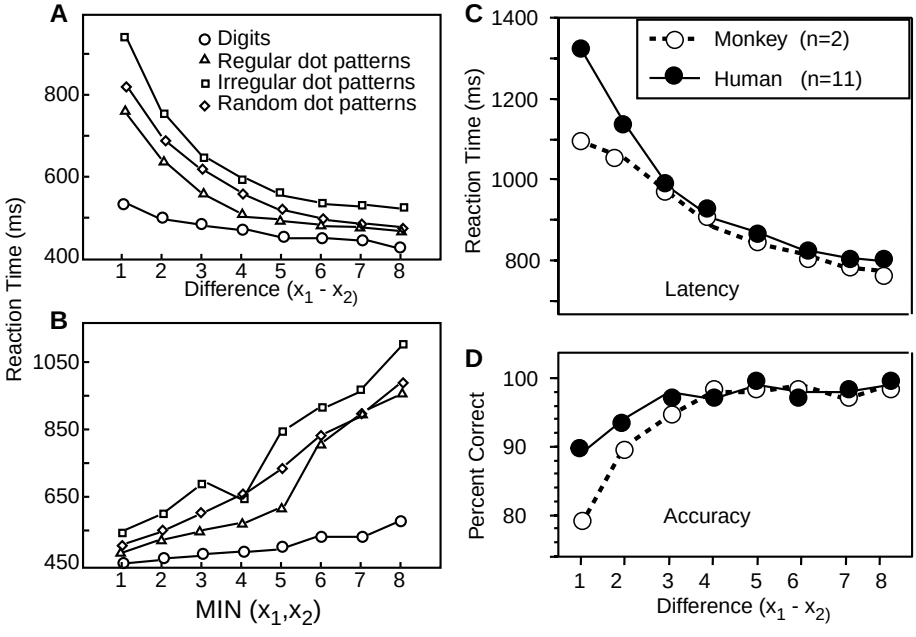


Fig. 1. The symbolic and nonsymbolic size and distance effects in the judgment of numerical order. (**A**) Time taken to make the order judgment as a function of the difference between two single digits (open circles) or between instantiations of the two numerosities (other symbols, nonsymbolic). (**B**) Reaction time as a function of the size of the smaller digit or number of stimulus items. (**C** and **D**) The distance effect for instantiated numbers is the same in humans and monkeys. [(A) and (B) are based on figures 1 and 2 in (*19*); (C) and (D) are based on figure 26.5 in (*12*).]

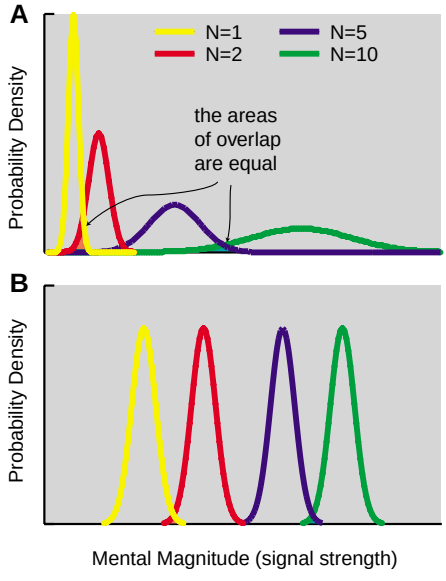


Fig. 2. Two common explanations for the size and distance effects. (**A**) Scalar variability. The mean mental magnitude is proportional to the number, as is the variability about this mean. Thus, the distributions are scale invariant, which means that the overlap between any two of them is determined by the ratio of their means. (**B**) Logarithmic compression. The mean mental magnitude is proportional to the logarithm of the number, whereas the variability is independent of it. Again, distributions for objective magnitudes that differ by a given ratio (e.g., 2:1) show the same overlap and, hence, the same potential for confusion about which distribution a particular signal properly belongs to.

merical reference may play a role in the emergence of a fully formed conception of number. The challenge now is to delineate that role.

References and Notes

1. L. Gleitman, A. Papafragou, in *Handbook of Thinking and Reasoning*, K. J. Holyoak, R. Morrison, Eds. (Cambridge Univ. Press, New York, in press).
2. D. Gentner, S. Golden-Meadow, Eds., *Language and Mind: Advances in the Study of Language and Thought* (MIT Press, Cambridge, MA, 2003).
3. S. C. Levinson, in *Language and Space*, P. Bloom, M. Peterson, L. Nadel, M. Garrett, Eds. (MIT Press, Cambridge, MA, 1996), Chap. 4.
4. R. Gelman, S. A. Cordes, in *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*, E. Dupoux, Ed. (MIT Press, Cambridge, MA, 2001), pp. 279–301.
5. B. Butterworth, *The Mathematical Brain* (McMillan, London, 1999).
6. C. R. Gallistel, *The Organization of Learning* (Bradford Books/MIT Press, Cambridge, MA, 1990).
7. J. A. Fodor, *The Language of Thought* (T. Y. Crowell, New York, 1975).
8. P. Gordon, *Science* **306**, 496 (2004).
9. P. Pica, C. Lemer, V. Izard, S. Dehaene, *Science* **306**, 499 (2004).
10. D. L. Everett, (2004) http://lings.ln.man.ac.uk/info/staff/DE/cultgram.pdf (cited by permission).
11. C. R. Gallistel, R. Gelman, in *Handbook of Thinking and Reasoning*, K. J. Holyoak, R. Morrison, Eds. (Cambridge University Press, New York, in press).
12. E. M. Brannon, H. S. Terrace, in *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*, M. Bekoff, C. Allen, Eds. (MIT Press, Cambridge, MA, 2002), pp. 197–204.
13. S. Dehaene, *The Number Sense* (Oxford University Press, Oxford, 1997).
14. R. Gelman, B. Butterworth, *Trends Cognit. Sci.*, in press.
15. L. Gleitman, J. Trueswell, K. Cassidy, R. Nappa, A. Papafragou, *Lang. Learn. Dev.*, in press.
16. S. Carey, *Daedalus* **133**, 59 (2004).
17. E. von Glaserfield, in *The Development of Numerical Competence: Animal and Human Models,* S. T. Boysen, E. J. Capaldi (Lawrence Erlbaum Associates, Hillsdale, NJ, 1993), pp. 225-244.
18. H. Davis, R. Pérusse, *Behav. Brain Sci.* **11**, 561 (1988).
19. P. B. Buckley, C. B. Gillman, *J. Exp. Psychol.* **103**, 1131 (1974).

REVIEW

The Role of the Medial Frontal Cortex in Cognitive Control

K. Richard Ridderinkhof,[1,2]* Markus Ullsperger,[3] Eveline A. Crone,[4] Sander Nieuwenhuis[5]

Adaptive goal-directed behavior involves monitoring of ongoing actions and performance outcomes, and subsequent adjustments of behavior and learning. We evaluate new findings in cognitive neuroscience concerning cortical interactions that subserve the recruitment and implementation of such cognitive control. A review of primate and human studies, along with a meta-analysis of the human functional neuroimaging literature, suggest that the detection of unfavorable outcomes, response errors, response conflict, and decision uncertainty elicits largely overlapping clusters of activation foci in an extensive part of the posterior medial frontal cortex (pMFC). A direct link is delineated between activity in this area and subsequent adjustments in performance. Emerging evidence points to functional interactions between the pMFC and the lateral prefrontal cortex (LPFC), so that monitoring-related pMFC activity serves as a signal that engages regulatory processes in the LPFC to implement performance adjustments.

Flexible goal-directed behavior requires an adaptive cognitive control system for selecting contextually relevant information and for organizing and optimizing information processing. Such adaptive control is effortful, and therefore it may not be efficient to maintain high levels of control at all times. Here we review recent studies in cognitive neuroscience that have advanced our understanding of how the brain determines and communicates the need to recruit cognitive control. Convergent evidence suggests that the posterior medial frontal cortex (pMFC) and lateral prefrontal cortex (LPFC) are important contributors to cognitive control. Our focus is on the role of the pMFC in performance monitoring, especially in situations in which pMFC activity is followed by performance adjustments. Evaluating the adequacy and success of performance is instrumental in determining and implementing appropriate behavioral adjustments. For instance, detection of a performance error may be used to shift performance strategy to a more conservative speed/accuracy balance. Based on the evidence reviewed below, we develop the tentative hypothesis that one unified function of the pMFC is performance monitoring in relation to anticipated rewards. The monitored signals may index the failure (errors or negative feedback) or reduced probability (conflicts or decision uncertainty) of obtaining such rewards, and as such signal the need for increased control.

[1]Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, Netherlands. [2]Department of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, Netherlands. [3]Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1A, 04103 Leipzig, Germany. [4]Center for Mind and Brain, University of California Davis, 202 Cousteau Place, Suite 201, Davis, CA 95616, USA. [5]Department of Cognitive Psychology, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, Netherlands.

*To whom correspondence should be addressed. E-mail: K.R.Ridderinkhof@uva.nl

Performance Monitoring

Flexible adjustments of behavior and reward-based association learning require the continuous assessment of ongoing actions and the outcomes of these actions. The ability to monitor and compare actual performance with internal goals and standards is critical for optimizing behavior. We first review evidence from primate, electrophysiological, and functional neuroimaging studies that points toward the importance of pMFC areas (Fig. 1A) in monitoring unfavorable performance outcomes, response errors, and response conflicts, respectively. These conditions have in common that they signal that goals may not be achieved or rewards may not be obtained unless the level of cognitive control is subsequently increased.

Although the pMFC can also be activated by positive events (such as rewards) (*1*, *2*), we focus here on negative events and their consequences. Because errors and conflicts are intrinsically negative, and because unfavorable outcomes are typically more consequential for the regulation of cognitive control than are favorable outcomes, our review focuses on the role of the pMFC in monitoring negative events.

Monitoring unfavorable outcomes. Electrophysiological recordings in nonhuman primates implicate the pMFC in monitoring performance outcomes. Distinct neuron populations in the pMFC, particularly in the supplementary eye fields and the rostral cingulate motor area (CMAr), are sensitive to reward expectancy and reward delivery (*1*, *3*, *4*). In addition, CMAr neurons exhibit sensitivity to unexpected reductions in reward (*5*). Likewise, specific groups of neurons in the depth of the cingulate sulcus (area 24c) react to response errors and to unexpected omissions of rewards (*5*). These findings are consistent with a role for these neuronal populations in comparing expected and actual outcomes.

Human neuroimaging studies implicate the pMFC, including the dorsal anterior cingulate cortex (ACC), along with other brain structures, in differential processing of unfavorable outcomes (Fig. 1B). These include studies using monetary rewards and punishments (*6*) and studies using abstract performance feedback (*7*). Similar parts of the pMFC are activated by primary reinforcers such as pain affect and pleasant tastes, suggesting that the pMFC plays a general role in coding the motivational value of external events.

Electrophysiological recordings in humans have identified the purported event-related brain potential correlate of the pMFC response to unfavorable outcomes: the feedback-related error-related negativity (or "feedback ERN"). This negative-polarity voltage deflection peaks approximately 250 to 300 ms after a stimulus indicating the outcome, and is greater in amplitude for negative performance feedback and outcomes indicating monetary losses than for positive feedback and monetary gains (*8*). The timing of this brain potential suggests that the pMFC computes or has access to a rapid evaluation of the outcome stimulus. Furthermore, initial studies report that the amplitude of the feedback ERN shows a graded sensitivity to the value of outcome stimuli that is normalized with respect to the subjectively expected outcome value (mean) and experienced range of outcome values (variance) (*9*).

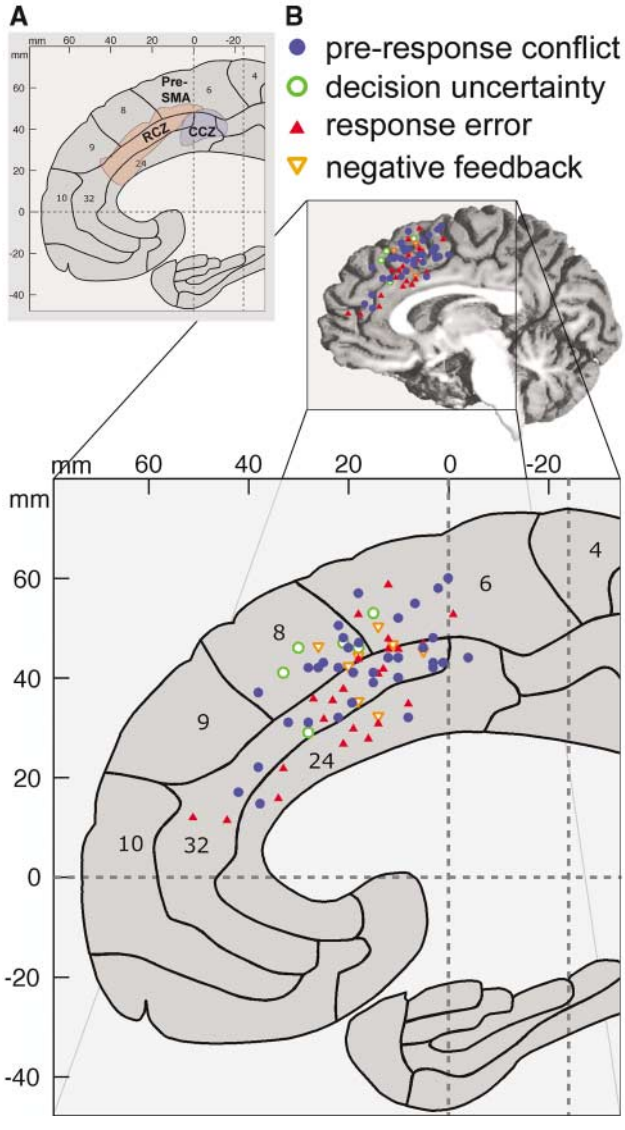


Fig. 1. Areas in the medial frontal cortex involved in performance monitoring. (**A**) Anatomical map of the medial frontal cortex. This is a schematic map of anatomical areas in the human pMFC, based on the atlas by Talairach and Tournoux (see supporting online material). The numbers indicate Brodmann areas. The area shaded in red encompasses the RCZ, and the area shaded in blue indicates the caudal cingulate zone (CCZ), as suggested by Picard and Strick (*11*). (**B**) Outcome of a meta-analysis of midline foci of activation reported in 38 fMRI studies published between 1997 and 2004 investigating brain activity associated with pre-response conflict, decision uncertainty, response errors, and negative feedback (*20*). In the upper part of the figure, the activation foci are superimposed on a saggital slice of an anatomical MRI scan at $x = 4$. In the lower part, the activation foci are superimposed on the enlarged schematic area map. The majority of activations cluster in the posterodorsal medial frontal cortex, in the region where areas 8, 6, 32, and 24 border each other.

Monitoring response errors. Primate studies show that, in addition to feedback-sensitive cells, the CMAr also contains error-sensitive cells (*4*, *10*). Corroborating these results, subsequent human functional neuroimaging studies have reported increased pMFC activation in response to errors as compared to correct responses in various two-alternative forced-choice tasks (*11*). The reported error-related activations cover a wide range along the anterior-posterior extent of the pMFC, with particular clustering in the rostral cingulate zone (RCZ) (*12*), the human homolog of the monkey's CMAr (Fig. 1B).

Consistent with these single-cell recordings and brain imaging studies, electrophysiological scalp recordings have found an error-sensitive event-related brain potential localized to the pMFC, which is attenuated in patients with damage to the dorsal ACC (*13*). This response-related ERN (or "response ERN") develops at the time of the first incorrect muscle activity and peaks about 100 ms later, indicating that the underlying generator has access to an efference copy of the initiated incorrect response (*14*). The response ERN is triggered by errors elicited under speeded response conditions, independent of the response effector (such as hands, feet, eyes, or voice), and increases in amplitude with the size or degree of error (*15*). Errors in these tasks result predominantly from premature responding, but continued stimulus processing after the response can provide sufficient information for outcome assessment. The morphology, polarity, and scalp distribution of the response ERN are similar to those of the feedback ERN, suggesting that the two ERN potentials may index a generic error-processing system in the pMFC.

A recent theory has extended the notion that the role of the dorsal ACC in coding outcome- and error-related information may be understood in terms of a common functional and neurobiological mechanism (*8*). The theory is predicated on prior research indicating that errors in reward prediction are coded by phasic changes in the activity of the midbrain dopamine system: a phasic increase when ongoing events are suddenly better than expected, and a phasic decrease when ongoing events are suddenly worse than expected (*16*). The theory builds on this research by proposing that these phasic dopamine signals are conveyed to the RCZ, where the signals are used to improve task performance in accordance with the principles of reinforcement learning. Furthermore, it proposes that the phasic dopamine signals modulate the activity of motor neurons in the RCZ, which is measurable at the scalp as changes in ERN amplitude. Phasic decreases in dopamine activity (indicating a negative reward prediction error) are associated with large ERNs and phasic increases (indicating a positive reward prediction error) with small ERNs.

A strong prediction of this theory is that the same region of the dorsal ACC should be activated by response errors and unexpected negative feedback. Also, during reward-based action learning, neural activity in this area should gradually propagate back from the feedback to the action that comes to predict the value of the feedback. These predictions have been confirmed using neuroimaging, ERN measurements, and computational modeling (*8*, *17*).

Monitoring response conflict. An alternative theory is that the pMFC, and in

particular the dorsal ACC, is involved in the monitoring of response conflict (*18*). Response conflict occurs when a task concurrently activates more than one response tendency; for example, when the stimulus primes a prepotent but incorrect response or when the correct response is underdetermined. Often, incorrect response tendencies are overridden in time by the overt correct response, resulting in high response conflict before the correct response (pre-response conflict). In contrast, occasional errors resulting from premature responding are characterized by response conflict after the response: The correct response tendency resulting from continued stimulus processing conflicts with the already executed incorrect response. In underdetermined responding (that is, under conditions requiring choosing from a set of responses, none of which is more compelling than the others), decision uncertainty occurs. Thus, decision uncertainty involves conflict similar to response conflict observed in tasks in which a prepotent response is overridden (*18*).

The conflict-monitoring theory is consistent with the neuroimaging evidence for pMFC activation in response to errors, reviewed above, and with the timing of the response ERN, indicating post-response conflict. In addition, the theory predicts that the pMFC should be active in correct trials characterized by high pre-response conflict, a prediction that has been confirmed by a large number of studies (Fig. 1B). Moreover, the predicted timing of such conflict-related activity is consistent with the occurrence of an ERN-like component, the N2, just before the response (*19*). Finally, the detection of high post-response conflict may be used as a reliable basis for internal error detection, thereby obviating the need for an explicit error detection mechanism (*19*).

The theory further holds that, upon the detection of response conflict, the pMFC signals other brain structures that the level of cognitive control needs to be increased.

Convergence and divergence in performance monitoring. The findings reviewed above suggest that the detection of unfavorable outcomes, response errors, response conflict, and decision uncertainty elicits largely overlapping clusters of activation foci in the pMFC. This assumption is consistent with a meta-analysis of the human neuroimaging literature (table S1), focusing on pMFC activations in response to these types of events (Fig. 1B) (*20*). The high degree of overlap should not be taken, however, as direct evidence for a generic role of neurons (or neuronal populations) in this brain area in monitoring various aspects of performance. First, although there is considerable overlap, there are some apparent differences as well, with foci associated with pre-response conflict clustering slightly more dorsally than foci activated during error and feedback monitoring (*21*, *22*). Second, single-cell recordings in monkeys suggest that different (neighboring) neurons within specific pMFC regions can be involved in different aspects of performance monitoring (*4*). Thus, the overlap between the activation foci identified in human neuroimaging studies does not necessarily imply identical functions for all neurons or neuronal ensembles within the pMFC.

A potential link between the outlined theories of pMFC functions is that pre-response conflict and decision uncertainty signal a reduced probability of obtaining reward, whereas errors and unexpected negative feedback signal the loss of anticipated reward. The pMFC, particularly the RCZ, is engaged when the need for adjustments to achieve action goals becomes evident. Interestingly, the monitoring processes examined here cluster primarily in the transition zone between the cingulate and paracingulate (areas 24 and 32), association (area 8), and premotor cortices (area 6), an area that has extensive connections with brain areas involved in the control of cognitive and motor processes and has been implicated in the regulation of autonomic arousal (*23*, *24*). This presumably places the pMFC in a strategically located position for signaling the need for performance adjustments and for interacting with brain areas involved in motor and cognitive, as well as autonomic and motivational, functions.

Performance Adjustments

Although the pMFC is consistently implicated in action monitoring, the mechanisms underlying the implementation of subsequent performance adjustments are less well understood. Two important questions are: (i) Is there a link between pMFC activation associated with performance monitoring and subsequent performance adjustments? (ii) What brain structures may be involved in the implementation of such control adjustments? In neuroimaging and neuropsychological studies, the LPFC has been broadly implicated in the coordination of adaptive goal-directed behavior (*25*–*29*). We review studies that address the first question, and we briefly evaluate the scant literature on functional interactions between the pMFC and LPFC in the service of adaptive control.

pMFC activity and immediate control adjustments. When stimuli elicit conflicting response tendencies or overt response errors, appropriate performance adjustments may be aimed not only at immediate correction of these tendencies but also at preventing errors on subsequent trials. A distinction can be made between two types of trial-to-trial performance adjustments: (i) shifts in the tradeoff between speed and accuracy of responding that place the cognitive system in a more cautious (as opposed to impulsive) response mode, and (ii) increases in control that improve the efficiency of information processing. Speed/accuracy tradeoffs may be expressed in ''post-error slowing,'' the observation that reaction times typically slow down after errors and correct, high-conflict trials (*18*). Changes in control, induced by such trials, can become evident in improved performance due to reduced interference from distracting information. For example, the increase in reaction times normally observed for incongruent stimuli (where target and distractor stimuli call for opposing responses) as compared to congruent stimuli (when distractors elicit the same action as the target stimulus) is typically reduced on trials after errors (*30*).

Several observations are consistent with a close link between modulations of pMFC activity and subsequent changes in performance. One study categorized trials in terms of their ERN amplitudes and found that the reaction time on the subsequent trial slowed progressively with increasing ERN amplitude on the current trial (*14*). In a similar vein, response errors on a two-alternative forced-choice task are foreshadowed by modulation of this pMFC activity during the immediately preceding (correct) response. Error-preceding trials were characterized by increased positivity in the time window typically associated with the ERN (*31*). This ''error-preceding positivity'' may reflect a transient disengagement of the monitoring system, resulting in occasional failures to implement appropriate control adjustments and hence in errors. Experimental factors that affect ERN amplitude may also affect subsequent performance adjustments. For example, alcohol consumption led to a reduction in the ERN amplitude and eliminated the post-error reduction of interference observed in a control condition (*30*). The relation between these findings and the associated neural circuitry was captured more directly in recent neuroimaging studies of Stroop task and response-inhibition performance (*32*, *33*): Post-hoc reaction time analyses revealed that greater ACC activity during error trials was associated with greater post-error slowing.

The latter studies also addressed the role of the LPFC in implementing control adjustments and its interaction with the pMFC. Trials exhibiting the greatest behavioral adjustments after errors and correct, high-conflict trials were associated with increased activity in the LPFC. Further, the degree of pMFC activity on conflict and error trials accurately predicted activity in the LPFC on the next trial. These and other findings are consistent with the idea that the pMFC, as a

monitor, and the LPFC, as a controller, interact in the regulation of goal-directed behavior (*18*).

pMFC activity and reward-based association learning. In addition to the link between pMFC activity and immediate adjustments in performance, there also seems to be a close relation between pMFC activity and reward-based association learning. A study of reward-based reversal learning in monkeys identified cells in the CMAr that fired only when two conditions were met: (i) reward was less than anticipated, and (ii) the reduction in reward was followed by changes in the monkeys' action selection (*5*). This finding has been corroborated by two recent functional magnetic resonance imaging (fMRI) studies of reversal learning, showing that ACC activity was observed under the same conjunctive condition (*34*, *35*). Reversal learning studies typically also show activation of the LPFC and other structures in association with changes in choice behavior (*36*). Whether these behavioral adjustments are implemented by or pMFC or whether the pMFC merely signals the LPFC or other structures to implement the adjustments remains to be explored.

Finally, there is evidence for an intimate relation between ERN amplitude and associative learning. In scalp electrophysiological activity, recorded from human participants who were required to learn stimulus-response contingencies on the basis of trial-to-trial positive or negative feedback, the feedback ERN to negative feedback decreased as participants were learning the contingencies, which is consistent with the theory discussed above that the ERN reflects a reward prediction error signal (*8*). Also, as participants learned the response associated with each stimulus, the response ERN associated with choice errors (provoked through the use of a stringent reaction time deadline) increased. In a temporal difference-learning model, not only did the ERN correlate with a reward prediction error but the brain activity underlying the ERN could also serve as a reinforcement learning signal for associative learning and hence optimizing task performance (*8*).

Conclusions and Future Directions

We have provided an overview of the evidence suggesting a critical role for the pMFC in performance monitoring and the implementation of associated adjustments in cognitive control. Our meta-analysis indicates that an extensive part of the pMFC—including areas 6, 8, 24, and 32, largely falling into a region referred to as the RCZ in humans—is consistently activated after the detection of response conflict, errors, and unfavorable outcomes. The similarities between two brain potentials generated by this area, the ERN and feedback ERN, are consistent with the view that the pMFC accommodates a unified functional and neurobiological performance-monitoring mechanism (*8*). This mechanism allows the pMFC to signal the likelihood of obtaining an anticipated reward (either definitive, as observed in studies of error detection and feedback processing, or probabilistic, as observed in studies of decision uncertainty and pre-response conflict).

Three conclusions from the meta-analysis should be emphasized. First, performance monitoring is associated with pMFC activations in a functionally integrated region (the RCZ) that cuts across various Brodmann areas beyond the "traditionally" reported ACC. Second, the most pronounced cluster of activations is in area 32 for all types of monitored events, suggesting the importance of this area for a unified performance monitoring function. Thus, the conclusion that error monitoring and conflict monitoring are performed by different areas, as derived from initial studies that were designed to identify differential involvement, is not ubiquitously confirmed by the meta-analysis. Third, activations related to pre-response conflict and uncertainty occur more often in area 8 and less often in area 24 than do activations associated with errors and negative feedback. Thus, although there is considerable overlap, there are some apparent differences as well, with activation foci associated with reduced probabilities of obtaining reward clustering slightly more dorsally than foci associated with errors and failures to obtain anticipated reward.

This generic monitoring function endows the pMFC with the capacity to signal the need for performance adjustment. Indeed, further evidence indicates a tight link between activity in this area and subsequent adjustments in performance, suggesting that the pMFC signals other brain regions that changes in cognitive control are needed. Although direct evidence is sparse, a likely candidate structure for effecting these control adjustments is the LPFC. Thus, monitoring-related pMFC activity may serve as a signal that engages control processes in the LPFC that are needed to regulate task performance in an adaptive fashion.

This conclusion notwithstanding, several questions remain. First, most studies of the pMFC and performance monitoring have tried to relate pMFC activity to control adjustments on the subsequent trial. An unresolved issue is whether the monitoring signal from the pMFC can also be used to resolve response conflicts on a within-trial basis (*34*). There is in principle no reason why such adjustments could not be implemented already within the same trial (to resolve conflict and correct the activation of inappropriate responses before they eventuate in an overt error). It is hard to tackle this question empirically using neuroimaging studies, because it requires disentangling the monitoring signal (indicating the need for control) and the answer to this signal (control implementation), which may be partly overlapping in time.

Another unresolved issue concerns the nature of the connection between the pMFC and LPFC. Anatomical studies in monkeys show dense reciprocal connections of the pMFC and LPFC (*37*, *38*). In humans, evidence for such connections is more indirect. Neuroimaging studies show concomitant activations in the LPFC and pMFC (*39*), suggesting close functional connectivity between these two areas. Little is known, however, about differential or selective reciprocal projections between various portions of the pMFC on the one hand and various subdivisions of the LPFC on the other. Possibly, this functional interplay is in part mediated by subcortical structures such as the basal ganglia and mesencephalic nuclei (*7*, *8*) or by the supplementary motor area (SMA) or pre-SMA (*29*, *40*).

Electrophysiological studies of patients with LPFC lesions have reported abnormal pMFC activity in response to errors (*41*). Such studies argue against the possibility of unidirectional information flow between the pMFC and LPFC, and instead suggest that performance monitoring and the regulation of cognitive control may be realized through intricate reciprocal projections between these two structures. It is a challenge for future research to further identify and characterize these interactions.

Although our review of the literature capitalizes on the role of the pMFC in performance monitoring, leading to performance adjustments on subsequent trials, other studies have suggested a more executive role for the pMFC in implementing control directly (*42*). Studies in nonhuman primates have shown that cells in the pMFC (especially in the monkey homolog of the RCZ) are well situated for this role, because this area has direct and indirect projections to primary and supplementary motor areas (*43*, *44*). It has been argued that some of these cells are involved in "goal-based action selection" (that is, selecting between competing actions in view of the anticipated reward associated with each of these actions) (*43*, *44*). The relation between these complementary functions remains to be further explored.

References and Notes

1. M. Shidara, B. Richmond, *Science* **296**, 1709 (2002).
2. B. Knutson, G. W. Fong, C. M. Adams, J. L. Varner, D. Hommer, *Neuroreport* **12**, 3683 (2001).
3. V. Stuphorn, T. L. Taylor, J. D. Schall, *Nature* **408**, 857 (2000).

4. S. Ito, V. Stuphorn, J. W. Brown, J. D. Schall, *Science* **302**, 120 (2003).
5. K. Shima, J. Tanji, *Science* **282**, 1335 (1998).
6. J. O'Doherty, M. L. Kringelbach, E. T. Rolls, J. Hornak, C. Andrews, *Nature Neurosci.* **4**, 95 (2001).
7. M. Ullsperger, D. Y. von Cramon, *J. Neurosci.* **23**, 4308 (2003).
8. C. B. Holroyd, M. G. H. Coles, *Psychol. Rev.* **109**, 679 (2002).
9. C. B. Holroyd, J. T. Larsen, J. D. Cohen, *Psychophysiology* **41**, 245 (2004).
10. H. Gemba, K. Sasaki, V. B. Brooks, *Neurosci. Lett.* **70**, 223 (1986).
11. M. Ullsperger, D. Y. Von Cramon, *Cortex*, in press.
12. N. Picard, P. L. Strick, *Cereb. Cortex* **6**, 342 (1996).
13. C. B. Holroyd, S. Nieuwenhuis, R. B. Mars, M. G. H. Coles, in *Cognitive Neuroscience of Attention*, M. I. Posner, Ed. (Guilford, New York, in press).
14. W. J. Gehring, B. Goss, M. G. H. Coles, D. E. Meyer, E. Donchin, *Psychol. Sci.* **4**, 385 (1993).
15. M. Falkenstein, J. Hoormann, S. Christ, J. Hohnsbein, *Biol. Psychol.* **51**, 87 (2000).
16. W. Schultz, *Neuron* **36**, 241 (2002).
17. C. B. Holroyd *et al.*, *Nature Neurosci.* **7**, 497 (2004).
18. M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, J. D. Cohen, *Psychol. Rev.* **108**, 624 (2001).
19. N. Yeung, M. M. Botvinick, J. D. Cohen, *Psychol. Rev.*, in press.
20. Materials and methods are available as supporting material on *Science* Online.
21. R. Hester, C. Fassbender, H. Garavan, *Cereb. Cortex* **14**, 986 (2004).
22. The majority of activations fall into the border zone between areas 8, 6, and 32, with some extension into area 24. Recent research in nonhuman primates seems to suggest a functional-anatomical dissociation of regions subserving pre-response conflict monitoring from structures sensitive to errors and omission of reward (*1*, *4*). Although in humans this view is still under debate (*11*, *13*, *21*), the present meta-analysis does not provide unequivocal evidence for or against such a dissociation. Activations related to pre-response conflict and uncertainty occur more often in area 8 and less often in area 24 than do signal increases associated with errors and negative feedback (area 8, 32.5% versus 9.7%; area 24, 7.5% versus 25.8%), supporting the dissociation view. However, both groups of activations cluster primarily in area 32 (pre-response, 42.5%; error, 41.9%), suggesting that pre- as well as post-response monitoring processes share at least one underlying structure. It seems that the currently available spatial resolution in fMRI, in conjunction with anatomical variability and differences in scanning and preprocessing methods between studies, limit the ability to resolve this debate about a possible dissociation in the range of 10 mm or less.
23. T. Paus, *Nature Rev. Neurosci.* **2**, 417 (2001).
24. H. D. Critchley *et al.*, *Brain* **126**, 2139 (2003).
25. E. K. Miller, J. D. Cohen, *Annu. Rev. Neurosci.* **24**, 167 (2001).
26. A. R. Aron, T. W. Robbins, R. A. Poldrack, *Trends Cogn. Sci.* **8**, 170 (2004).
27. D. Badre, A. D. Wagner, *Neuron* **41**, 473 (2004).
28. S. A. Bunge, K. N. Ochsner, J. E. Desmond, G. H. Glover, J. D. E. Gabrieli, *Brain* **124**, 2074 (2001).
29. M. Brass, D. Y. von Cramon, *J. Cogn. Neurosci.* **16**, 609 (2004).
30. K. R. Ridderinkhof *et al.*, *Science* **298**, 2209 (2002).
31. K. R. Ridderinkhof, S. Nieuwenhuis, T. R. Bashore, *Neurosci. Lett.* **348**, 1 (2003).
32. J. G. Kerns *et al.*, *Science* **303**, 1023 (2004).
33. H. Garavan, T. J. Ross, K. Murphy, R. A. Roche, E. A. Stein, *Neuroimage* **17**, 1820 (2002).
34. G. Bush *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 523 (2002).
35. J. O'Doherty, H. Critchley, R. Deichmann, R. J. Dolan, *J. Neurosci.* **23**, 7931 (2003).
36. R. Cools, L. Clark, A. M. Owen, T. W. Robbins, *J. Neurosci.* **22**, 4563 (2002).
37. J. F. Bates, P. S. Goldman-Rakic, *J. Comp. Neurol.* **336**, 211 (1993).
38. M. Petrides, D. N. Pandya, *Eur. J. Neurosci.* **11**, 1011 (1999).
39. L. Koski, T. Paus, *Exp. Brain Res.* **133**, 55 (2000).
40. K. Fiehler, M. Ullsperger, D. Y. von Cramon, *Eur. J. Neurosci.* **19**, 3081 (2004).
41. W. J. Gehring, R. T. Knight, *Nature Neurosci.* **3**, 516 (2000).
42. M. I. Posner, G. J. DiGirolamo, in *The Attentive Brain*, R. Parasuraman, Ed. (MIT Press, Cambridge, MA, 1998), pp. 401–423.
43. K. Matsumoto, K. Tanaka, *Science* **303**, 969 (2004).
44. K. Matsumoto, K. Tanaka, *Curr. Opin. Neurobiol.* **14**, 178 (2004).
45. This research was supported by a TALENT grant (E.A.C.) and a VENI grant (S.N.) of the Netherlands Organization for Scientific Research and by the Priority Program Executive Functions of the German Research Foundation (M.U.). Helpful comments by S. Bunge are gratefully acknowledged.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5695/443/DC1
Materials and Methods
Table S1
References

REVIEW

Neuroeconomics: The Consilience of Brain and Decision

Paul W. Glimcher[1]* and Aldo Rustichini[2]

Economics, psychology, and neuroscience are converging today into a single, unified discipline with the ultimate aim of providing a single, general theory of human behavior. This is the emerging field of neuroeconomics in which consilience, the accordance of two or more inductions drawn from different groups of phenomena, seems to be operating. Economists and psychologists are providing rich conceptual tools for understanding and modeling behavior, while neurobiologists provide tools for the study of mechanism. The goal of this discipline is thus to understand the processes that connect sensation and action by revealing the neurobiological mechanisms by which decisions are made. This review describes recent developments in neuroeconomics from both behavioral and biological perspectives.

The full understanding of utility will come from biology and psychology by reduction to the elements of human behavior followed by a bottom-up synthesis, not from the social sciences by top-down inference and guesswork based on intuitive knowledge. It is in biology and psychology that economists and social scientists will find the premises needed to fashion more predictive models, just as it was in physics and chemistry that researchers found the premises that upgraded biology. (p. 206) (*1*)

Consider the famous St. Petersburg paradox (*2*). Which of the following would you prefer, $40 or a lottery ticket that pays according to the outcomes of one or more fair coin tosses: heads you get $2 and the game ends, tails you get another toss and the game repeats, but now if the second toss lands heads up you get $4, and so on. If the nth toss is the first to land heads up, you get 2^n dollars. The game continues, however long it takes, until the coin lands heads up. We can assess the average objective, or expected, value of this lottery by multiplying the probability of a win on each flip by the amount of that win:

$$\text{Expected value} = (0.5 \times 2) + (0.25 \times 4) + (0.125 \times 8)\ldots$$
$$= 1 + 1 + 1 + \ldots$$

This simple calculation reveals that the expected value of the lottery is infinite even though the average person is willing to pay less than $40 to play it. How could this be?

For an economist, any useful explanation must begin with a set of assumptions that renders behavior formally tractable to coherent theoretical and mathematical analysis. Economists therefore explain this behavior by assuming that the desirability of money does not increase linearly, but rather grows more and more slowly as the total amount at stake increases. For example, the desirability of a given amount might be a power function

[1]Center for Neural Science, New York University, New York, NY 10003, USA. [2]Department of Economics, University of Minnesota, Minneapolis, MN 55455, USA.

*To whom correspondence should be addressed. E-mail: glimcher@cns.nyu.edu

of that amount, as shown by the black line in Fig. 1. A decision-maker for whom the subjective value, or utility, of money grew in this fashion would then determine the desirability, or expected utility, of the St. Petersburg lottery by multiplying the probability of a win on each flip by the utility of the amount won on that flip, and thus he might well be willing to pay less than $40 to play this game.

From the point of view of a psychologist attempting to understand and explain this same phenomenon, it is the nature of risk aversion and the psychological mechanisms that this set of preferences reveals that become the subject of explanatory study. The psychological mechanism that accounts for risk aversion in human subjects, for example, has been shown to be more sensitive to monetary losses than to monetary gains. Further, psychologists have suggested that subjective utilities are computed with regard to somewhat arbitrary and idiosyncratic monetary reference points, or frames, set by yet other psychological processes (*3*). Psychologists use observations like these to argue that human choosers are endowed with a particularly strong fear of losses and that they weigh the merits of all possible gains and losses relative to a psychological benchmark: The psychological approach seeks empirically to describe minimally complex behavioral tendencies, modules, or heuristics that can account for the actions of human choosers.

A traditional neurobiological perspective uses yet another approach: A hungry bird is shown a tray that contains five millet seeds and repeatedly permitted to fly to the tray and eat the seeds. At a neurobiological level, the study of this behavior begins with the assumption that the visual stimulus of the five seeds must somehow propagate through the sensory system of the animal to trigger activation in orienting circuits that move the bird to the seeds. Next, the same bird is permitted to fly to a second tray covered by a piece of paper. When the bird displaces the cover, half of the time it reveals 12 seeds and half of the time it reveals nothing. Mechanistically, the visual stimulus must again trigger an orienting response, and presumably in this case the strength with which visual signals connect synaptically to the orienting circuits reflects both the number of seeds that the bird earns and the likelihood that seeds will be found under the paper. Lastly, both trays are presented, and the bird is observed to fly toward the tray that may contain 12 millet seeds. A standard neurobiological explanation (*4*, *5*) presumes that under these circumstances the two different behavioral circuits compete. In this case the synapses that elicit an orienting response to the covered tray are stronger and thus control behavior. The neurobiological explanation specifies the minimal neural circuitry required to account for the observed behavior of the bird.

What is striking about explanations of choice behavior by economists, psychologists, and neurobiologists is the different levels at which they operate. The economic approach attempts to describe globally all choice behavior with a single logically consistent formalism. The psychological approach examines the ways in which subjective and objective estimates of value differ and posits psychological modules that can account for these observed behavioral preferences. The neurobiological explanation starts with the simplest possible neural circuitry that can account for the simplest measurable elements of behavior. It seems obvious that these different levels of explanation should be linked, but how can such a linkage be accomplished? We argue that a unified explanation of decision-making is not only possible but has recently begun and that, when the linkage between these three levels of explanation has matured, a new, more powerful decision science rooted in a neuroeconomic approach will have been developed.

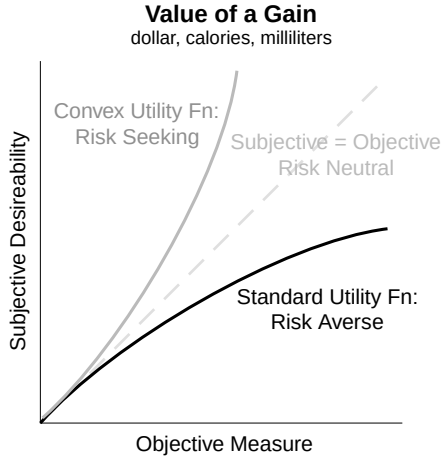


Fig. 1. Bernoulli's notion of subjective value or utility. The black line plots the typical relationship between objective and subjective valuations of an action. As the objective value of a gain increases, the subjective desirability, or utility, grows more slowly. Bernoulli demonstrated that this relationship could account for the observation that humans are typically risk-averse. The solid gray line plots a condition in which subjective value grows more quickly than objective value, a preference structure that would yield risk-seeking behavior.

A second claim we will make is that once this reconstruction of decision science is completed, many of the most puzzling aspects of human behavior, aspects that economic theory, psychological analysis, or neurobiological deconstruction have failed to explain, will become formally and mechanistically explicable. The claim is, in essence, that a decision science that simultaneously engaged all three approaches would be more heavily constrained and at the same time would have much greater explanatory power than do any of these three approaches operating alone. We will see examples of how this synthetic approach would operate in principle and early attempts at synthetic solutions below.

This reconstruction of the study of decision is also going to be the appropriate basis for a more ambitious theory that explains not just how we make decisions but why. That such an explanation is necessary and possible is indicated by the fact that fundamental features of decision making are common to many species. For example, risk aversion as shown by the St. Petersburg paradox has been described in many species. Studies of birds making choices in risky environments produce a behavior best described by a utility function (Fig. 1) (*6*, *7*). We know that humans and birds deviated from a common reptilian ancestor at least 200 million years ago, but this basic function for choice has remained essentially unchanged. Such commonalities make a clear suggestion: A utility function of this type probably is an efficient and evolved feature of vertebrate choice. For example, Robson (*8*) provides a justification of why a utility function might be an evolutionary optimal response to changing environments. Just as information theory was used by Barlow (*9*) to explain why animals as diverse as horseshoe crabs and cats use similar encoding schemes in their visual systems, an economic theory that relates utility to Darwinian fitness must serve as an overarching tool for understanding vertebrate choice behavior.

Linking the Decision Sciences

Subjective desirability. The central concept in modern economic theory is the notion of subjective utility: Preferences must be described as subjective properties of the chooser. Surprisingly, the notion that preferences are represented in the nervous system, that these preferences are subjective, and that they guide the production of action has only recently entered the neurobiological mainstream. We believe that this has been a critical flaw in neurobiological studies, because it is essential that economics, psychology, and neuroscience acknowledge a common phenomenological base to achieve a reductive unification of the decision sciences. The concepts that guide the behavioral study of decision-making must also guide the mechanistic study of that process.

In part, this preference-free approach may have arisen from neurobiology's roots in the stimulus-response physiology of the

early twentieth century (*10*). Working from Descartes' philosophy, Sherrington (*11*) proposed that physiologists should work to link stimulus and response directly through what Pavlov (*12*) would later call a "definite nervous path." Scientists like Sherrington and Pavlov proposed that it was the role of neuroscience to chart these stimulus-response connections through the nervous system.

A critical step beyond this initial framework was a recent effort to explain more complicated behaviors and to focus on actions for which deterministic sensory-to-movement mapping approaches were insufficient. Newsome and his colleagues (*13*, *14*) made that step in the late 1980s when they examined perceptual decision-making by monkeys viewing ambiguous sensory stimuli. In those experiments, monkeys stared at a display of chaotically moving spots of light. On training trials, a subset of the spots moved coherently in a single direction, whereas the remaining spots moved randomly (*15*). The direction of this coherent motion indicated which of two possible saccadic eye movements would yield a fruit juice reward, and at the end of each trial animals were free to make a saccade. If they made the correct movement, they then received the reward. On a critical subset of trials, however, monkeys viewed displays in which none of the dots moved in a single coherent direction, and thus the display provided no information from which the location of the rewarded eye movement could be deduced. Under these conditions, Newsome and his colleagues found that the firing rates of single neurons in the middle temporal visual area (area MT) were still correlated with the behavior of the animals, even when that behavior could not be predicted from the properties of the stimulus. Newsome and his colleague Shadlen's subsequent studies revealed the basic neurobiological substrate for perceptual decision-making and showed convincingly that this circuit could not be modeled simply as a single "definite nervous path" from stimulus to response (*16*).

This work, in turn, accelerated studies of the posterior parietal cortex, an area interposed between many of the sensory circuits and motor circuits of the primate brain, which appeared to play a critical role in the perceptual decision-making Newsome has studied (*17*, *18*). Platt and Glimcher (*19*) made an important advance when they extended Newsome's approach by proposing that posterior parietal cortex might play a role in decision-making in an economic sense and that it might encode the desirabilities of making particular movements.

In Platt and Glimcher's experiments, trained rhesus monkeys were allowed to participate in repeated rounds of a simple lottery while the activity of nerve cells in the posterior parietal cortex was monitored. At the beginning of each round, two yellow spots were illuminated on a screen, one to the left and one to the right of where the monkey was looking. This began the lottery phase of the round, a period during which the monkey did not know whether the left or right light would be offered as a prize at the end of that round. At the end of this phase, a third light changed color to red or green, indicating which of the two initial lights had been randomly selected to yield a fruit juice reward on that particular round. The monkey received the fruit juice if he oriented to the selected light at the end of the round. While monkeys played hundreds of rounds of this game, Platt and Glimcher systematically varied either the relative probabilities that the left or right lights would be selected at the end of each round or the size of the reward associated with each. These two variables were selected because economic theories assess desirability by combining the value and likelihood of gain in some subjective manner. Platt and Glimcher found that some parietal neurons did indeed encode the value and likelihood of reinforcement during the lottery phase of each round. Under these conditions, the brains of the monkeys explicitly encoded something very much like the economically defined expected value or expected utility of each light in this simple lottery task.

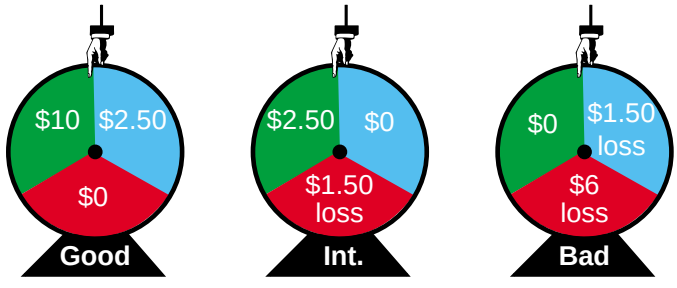


Fig. 2. The three lotteries used in the Breiter and colleagues experiment.

Subsequent studies of human decision-making using functional magnetic resonance imaging (fMRI) have yielded similar conclusions. Knutson and colleagues (*20*) have shown, for example, that activity in the human striatum is correlated with the magnitude of the monetary reward subjects earn during lotteries, and Paulus and colleagues (*21*) have shown a similar result in the human posterior parietal cortex. In a particularly interesting study, Breiter and colleagues (*22*) (Fig. 2) presented human subjects on sequential rounds with one of three possible lotteries. In lottery one (the good lottery), they faced equal chances of winning \$10, \$2.50, or \$0. In lottery two (the intermediate lottery), they faced equal chances of winning \$2.50, winning \$0, or losing \$1.50. In lottery three (the bad lottery), they faced equal chances of winning \$0, losing \$1.50, or losing \$6. At the beginning of each round, the subjects were told which lottery they would be playing, and the average activity in many brain areas was simultaneously measured. After that measurement was complete, the lottery was actually played and the humans were then told how much real money they had earned on that round. Importantly, all three of these lotteries present a one-third possibility of winning \$0, but they do so under different conditions. In the good lottery, winning \$0 is the worst possible outcome, whereas in the bad lottery it is the best. The psychologists Kahneman and Tversky (*23*) have shown that, when a human participates in this good lottery, they find winning \$0 to be an intensely negative outcome whereas when a human participates in the bad lottery, they find winning \$0 to be a positive outcome; subjective utilities are computed with regard to a reference frame. Breiter and colleagues found that the activity of the sublenticular extended amygdala encoded the desirability of each lottery, taking into account this behaviorally described framing effect.

Other recent neurobiological studies have revealed yet other neurally encoded variables that include the log likelihood that a given eye movement will result in a reward (*24*, *25*), the very closely related integral of perceptual signals indicating which saccade will be rewarded in the Newsome task (*26*, *27*), the average rate at which a saccade has been rewarded in the recent past (*28*), the instantaneous likelihood, or hazard, that a reinforced saccade will be instructed (*29*), and combinations of these variables (*30*).

Strategic thinking. All of these results suggest that classical utility theory can be used as a central concept for the study of choice in economics, psychology, and neuroscience. In the middle of the twentieth century, however, economists pushed utility theory beyond this boundary, enhancing it to include the study of the strategic interactions which arise when decision makers confront intelligent opponents. Extending the concept of subjective utility, VonNeumann and Morgenstern (*31*) and Nash (*32*, *33*) developed a formal utility-based economic approach in the theory of

games. Recently, Lee and his colleagues (*34*) and Dorris and Glimcher (*35*) have begun to link the neurobiological corpus to this literature by examining the activity of single neurons in awake-behaving monkeys engaged in decision-making during strategic conflicts. In Dorris and Glimcher's study, two opponents face each other, an employer and an employee. On each round of the game the employee must decide whether to go to work, in which case he earns a fixed wage, or whether to shirk, in hopes of earning his wage plus a bonus. The goal of the employee is simply to maximize his gains in terms of salary and bonus. The employer, on the other hand, must decide between trusting his employee to arrive for work or spending money to hire an inspector who can actually check and see whether the employee arrived for work that day. The goal of the employer is to spend as little as possible on inspections while maximizing the employee's incentive to work.

The inspection game is of particular interest to game theorists and economists because rational strategies for utility maximization during this strategic conflict lead to predictable outcomes according to an equilibrium theory originally developed by John Nash in the 1950s. Nash (*36*) equilibrium theory describes how, when the cost of inspection to the employer is set high, the efficient strategy for both players converges on a solution in which the employee manages to shirk fairly often. Conversely, a low inspection cost to the employer defines a theoretical equilibrium solution in which shirk rates are low. One of the fundamental insights this formal analysis reveals is that at a mixed strategy equilibrium, a situation in which a rational player should distribute his actions amongst two or more alternatives in an unpredictable fashion, the desirability of the two or more actions in equilibrium must be equivalent. The Nash approach argues, essentially, that a behavioral equilibrium occurs when the desirability of working and shirking are rendered equal by the behavior of one's opponent irrespective of how often that equilibrium requires that one work. When Dorris and Glimcher examined the activity of neurons in the posterior parietal cortex of monkeys playing the inspection game, they found neurons that carried a signal that behaved like relative expected utility. When the monkey's behavior was well predicted by the Nash equations, neural activity was fixed at a single level irrespective of the frequency with which the monkey chose to make a particular response, even though these same neurons were strongly modulated by changes in the value of actions during lottery tasks.

Research on human-human strategic interactions that are well described by classical game theory are also now under way in a number of laboratories (*37*). Like the earlier fMRI studies of simple decision-making tasks, these experiments are also beginning to shape the common ground between economics, psychology, and neuroscience. Taken together, these findings suggest that at least under some circumstances decisions may actually be made in the primate neuro-architecture in a manner long suspected by economists and now being actively analyzed by psychologists and neuroscientists: Neural circuits may compute and represent the desirability of making a response. Economics, psychology, and neuroscience do seem to be converging around a common conceptual framework. All three disciplines are beginning to acknowledge that decision-making involves the representation of subjective desirabilities. The challenge that this convergence around a single concept poses, however, is to leverage the intersection of these three disciplines to explain choice behavior that cannot be described with the common framework of utility theory; these are classes of behaviors which have stymied traditional economics and which have lain far beyond the reach of traditional neuroscience. If it is to be of value, the goal of a unified decision science will have to be to use all three sets of approaches simultaneously to gain traction in this new territory.

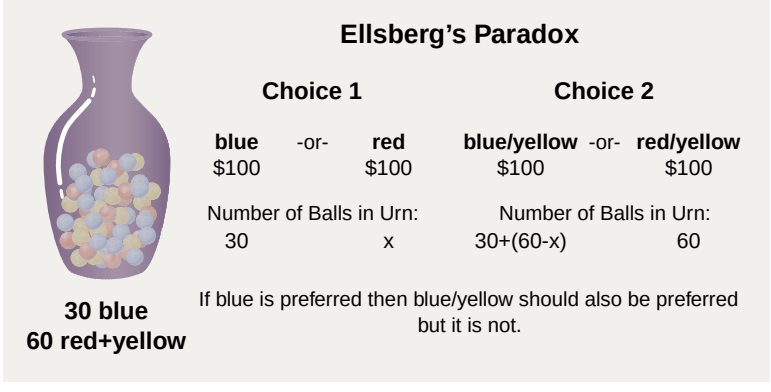


Fig. 3. Ellsberg's paradox. If blue is preferred in choice one, then blue/yellow should logically be preferred in choice two. Surprisingly, this is rarely the case.

Beyond Classical Concepts

Choice under risk. As we have seen, the introduction of the concept of expected utility solved the St. Petersburg's puzzle and formed the core of neoclassical economics. Subsequent puzzles and paradoxes, however, have plagued this solution. In Ellsberg's (1961) paradox (*38*) (Fig. 3) you are presented with an urn, and you are told that it contains 90 balls. Of these, 30 are blue, and 60 are either red or yellow; any proportion is possible. You are then offered a choice between a lottery that pays $100 if a blue ball is drawn (a 1/3 probability) and one that pays $100 if a red ball is drawn. The probability of a red draw is unspecified or ambiguous: It is a choice between an event with a known probability and an event with an unknown probability. Under these circumstances, people typically choose the first lottery, which wins if a blue ball is drawn. According to expected utility theory they could only do so if they believe that there are fewer than 30 red balls in the urn or, equivalently, that there are more than 30 yellow balls. Then (before any balls are actually drawn, but with the same urn standing in front of you) you are asked to choose again, this time between a lottery that pays $100 on either blue or yellow and one that pays $100 on either red or yellow. Now the likelihood of winning is clear in the second case (a 2/3 probability of winning $100) but unclear in the first case (a probability between 1/3 and 1). People this time typically choose the second lottery. The first lottery seems less attractive, because there might be too few yellow balls. Is there anything wrong with this behavior? If expected utility theory is correct, then there certainly is: You cannot think that there are too few and too many yellow balls in the urn at the same time.

The Ellsberg paradox is just one of many demonstrations presented in the last half of the twentieth century that were considered formal falsifications of expected utility theory. An even earlier example is Allais' paradox (*39*), based on the idea that a certain outcome may be perceived as more desirable, in a qualitatively different way, than any random outcome, even if very likely (*40*). These examples proved that expected utility theory as originally proposed could not be globally correct; at best it could only predict choices under some circumstances. This has led economists and social psychologists both to attempt modifications to expected utility theory and to replace it outright. Both the modifications and replacements have provided important and economically powerful insights into choice behavior but have not yet provided a global theory of

choice that can truly replace expected utility theory.

The emerging discipline of neuroeconomics offers a new strategy both for testing existing models of all types and for developing new models with empirical techniques. If we succeed in understanding mechanistically how choices that violate expected utility theory are made at a neural level, then a new global theory of choice will be developed. To that end, a number of laboratories are now beginning to reexamine the conditions under which expected utility theory fails.

The reason that expected utility theory fails under some conditions may be that choosers use more than one evaluative mechanism at a neurobiological level (*41*). For example, in Dickhaut *et al*. (*42*), the processes involved when a certain outcome is one of the options are different from those involved when only random outcomes are at stake, providing an explanation of the Allais' paradox cited above.

Under many conditions these mechanisms may work together to yield choices similar to those predicted by expected utility theory but may produce odd results when used in isolation, in novel combinations, or in situations for which they are ill suited. Recent work by Damasio and colleagues [for example, (*43*)] on the class of behavioral paradoxes from which the Ellsberg example is drawn seem to support this conclusion. These studies suggest that an ambiguity-sensitive mechanism associated with the expression of emotion may reside, at least in part, in the ventromedial prefrontal cortex (VMPFC) (Fig. 4) and may be responsible for choice under some but not all conditions. These researchers and others have shown that patients with damage to this area have an impaired ability to make some classes of decisions and have difficulties planning their work and choosing friends. Further, the actions these individuals do elect to pursue often lead to financial as well as personal losses. Yet despite these specific failures, patients with damage to the VMPFC show normal performance on multiple-choice tests of intelligence.

These observations and others like them have led Damasio to propose that the inability of patients with VMPFC lesions to make advantageous decisions under some circumstances is caused by damage to an emotional mechanism that stores and signals the value of future consequences of an action, the somatic marker hypothesis. The hypothesis proposes that, because they lack this emotional mechanism, the patients must rely on other brain mechanisms that achieve a different analysis of the numerous and often conflicting options involving both immediate and future consequences. This other mechanism, operating alone, is hypothesized to produce decisions that are less efficient and slower than those produced by a normal, intact, system.

The importance of the emotion-related VMPFC for regular decision-making has been confirmed by experiments where subjects were asked to make choices among a group of alternatives that carry a monetary reward (typically by selecting one card at a time from four different decks of cards), but for which the probability of reward is unspecified (*44*). This is precisely the ambiguous situation that produces Ellsberg's paradox. Under these conditions, patients with VMPFC lesions seem to lack an aversion to ambiguity or losses that normal subjects have, an aversion that may be quite advantageous under many conditions. Further support for this hypothesis comes from brain imaging studies. For example, O'Doherty *et al*. (*45*) have shown that the VMPFC is relatively more active when human subjects are actively learning about the availability of rewards and punishments during one of these ambiguous choice tasks.

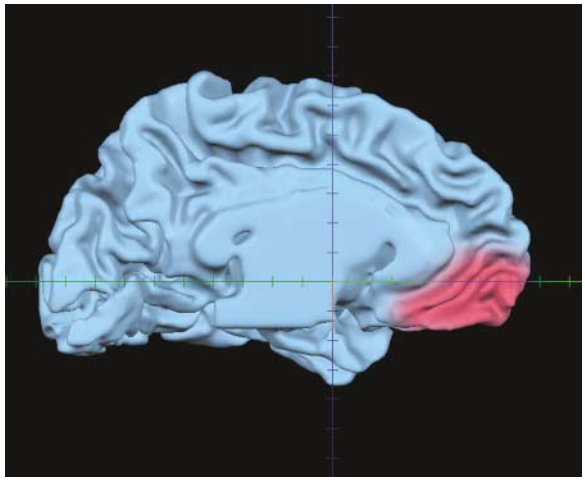


Fig. 4. Medial view of the left half of a human brain, with the front of the brain on the right side of the image. The human ventromedial prefrontal cortex is shown in red.

The process may be very different, however, when subjects simply choose between options without any feedback or learning taking place at the same time. For example, Rustichini *et al*. (*46*) asked normal subjects to make choices among ambiguous lotteries, risky lotteries, and certain outcomes while their brain activity was monitored. Subjects were paid for the outcome of their choices, but the outcome was communicated only after the experiment was over. Under these conditions, the VMPFC did not show any activation; it was actually less active when choices were being made than when subjects waited between trials. These results suggest that emotional circuits may be important in learning and processing information, rather than in selecting among alternatives.

Together, these data may begin to explain, in a mechanistic way, how information is analyzed when at least one class of behavior which is not predicted by the expected utility theory is produced. The process of learning and evaluating feedback may involve emotion-related areas. Ambiguity aversion, whether advantageous or disadvantageous in a particular situation, may become explicable as we learn more about the computations that brain areas like the VMPFC perform.

Strategic cooperation. As with the Ellsberg paradox, challenges have also been raised recently to classical game theory. In a path-breaking study, Guth *et al*. (*47*) analyzed the behavior of subjects playing the ultimatum game. In this game, a first player, the proposer, has $10 to split with a second player. He can offer any amount between zero and $10. The second player is informed of the offer and can accept or refuse. If she accepts, the split is made. If she refuses, both players get nothing. The prediction of a restrictive concept of game theory, the subgame perfect equilibrium, is that for any positive amount offered by the proposer, the second player knows that she faces a choice between gaining nothing (if she refuses the offer) or something (if she accepts). The proposer should therefore always offer the minimum possible split to player two, who should always accept. Contrary to this prediction, the robust experimental finding is that low offers (typically $2 or even $3) are consistently refused. The second player appears to prefer, under these conditions, to gain nothing. Anticipating this, proposers typically avoid low offers.

Although expected utility theorists have proposed some explanations for this behavior, it may well be that by analyzing the neural circuits active during the ultimatum game we may be able to both explain the causes of this behavior and to predict it. Studying the ultimatum game in subjects undergoing brain scans, Sanfey *et al*. (*48*) found that offers refused by the second players activated specific brain circuits in those players, and interestingly these brain circuits are also associated with emotional arousal: the anterior insula (AI, associated with disgust, both physical and emotional), the dorsolateral prefrontal

cortex (DLPFC, associated with goal maintenance and executive control), and the anterior cingulate cortex (ACC, associated with detection of cognitive conflict). Also significant is the correlation of activation with choices: An activation of the AI is positively correlated with rejection, suggesting that an emotional arousal associated with a low offer is correlated with rejection. The overall picture is that offers we might consider unfair may activate emotional circuits of the brain involved in the decision to reject an offer. If we can come to more fully understand how these circuits reach this conclusion, then a behavior that was difficult for classical game theory to predict may become fully explicable with the synthetic approach that neuroeconomics provides.

A similar line of investigation has examined interplayer cooperation during single rounds of the trust game. In this game, two players move sequentially. The first player can decide to transfer a sum of money out of an initial endowment that she receives into an investment pool that immediately triples in value. The second player then gains control of the investment and can divide it between the two players in any way he chooses. In this game, the only Nash equilibrium choice for the first player is to transfer nothing into the investment. Were she to make any transfer, the second player should take all of the money for himself. However, in real experiments the first player typically does transfer a significant amount into the investment, and the second player reciprocates by returning part of the pool. McCabe *et al*. (*49*) had subjects play the trust game both against a human opponent and against a computer program which, they were told, would play a human-like strategy. Under these conditions McCabe and colleagues found that subjects were more likely to cooperate with real humans than with computers and that cooperators have a significantly different brain activation in the two conditions. Cooperation is associated with activation of the anterior paracingulate cortex, a brain region associated with (*50*, *51*) interpreting and monitoring the mental state of others.

Although these studies are in their early stages, they suggest the existence of specific brain components that make specialized contributions to decision making. The challenge that these studies face is to derive detailed computational models of the neural mechanisms, which will make neuroeconomic models broadly predictive as well as explanatory.

Summary

Economics, psychology, and neuroscience are converging into a single, unified field aimed at providing a theory of human behavior. In this enterprise, the method and the standard set by neuroscience is the final goal: a reconstruction of the process and mechanism that goes from a stimulus presented to the subject to his final action in response. Economics provides the conceptual structure and the object of the analysis. In this emerging view, people are seen as deciding among options on the basis of the relative desirability of each option. This is true when they are in isolation as well as when they are in strategic (interaction with few persons) and market (interaction with a large number) environments. The recent research we have been surveying describes how desirability is realized as a concrete object, a neural signal in the human and animal brain, rather than as a purely theoretical construction. Desirability is computed and is represented in the brain, and we now have the means to test, measure, and represent this activation.

But the complete reconstruction of the decision process, and hence of human behavior, is not going to be easy, because two of the cornerstones of economic analysis, subjective utility theory and Nash equilibrium, provide, even from the descriptive point of view, an incomplete picture. For example, desirability as represented by the simple economic formalism of expected utility may be appropriate only in simple conditions, where ambiguity is excluded. A more general notion is needed and, as we have seen, is beginning to be investigated and developed by psychologists and economists working together. The goal of the emerging neuroeconomic program will have to be a mechanistic, behavioral, and mathematical explanation of choice that transcends the explanations available to neuroscientists, psychologists, and economists working alone. Although it is unclear today how complete this explanation will ultimately be, neuroeconomic approaches have already begun to yield substantial fruit and to fuse natural and social scientific approaches to the study of human behavior.

References and Notes

1. E. O. Wilson, *Consilience* (Knopf, New York, 1998).
2. D. Bernoulli, *Econometrica* **22**, 23 (1954).
3. D. Kahneman, P. Slovic, A. Tversky, Eds, *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge Univ. Press, Cambridge, 1982).
4. C. S. Sherrington, *The Integrative Action of the Nervous System* (Scribner, New York, 1906).
5. I. P. Pavlov, *Conditioned Reflexes an Investigation of the Physiological Activity of the Cerebral Cortex* (Oxford Univ. Press, Oxford, 1927).
6. T. Caraco, S. Martindale, T. Whittam, *Anim. Behav.* **28**, 820 (1980).
7. A. Kacelnik, M. Bateson, *Trends Cogn. Sci.* **1**, 304 (1997).
8. A. Robson, *J. Polit. Econ.* **109**, 900 (2001).
9. H. B. Barlow, in *Current Problems in Animal Behavior*, W. H. Thorpe, O. L. Zangwill, Eds. (Cambridge Univ. Press, Cambridge, 1961).
10. P. W. Glimcher, *Decisions, Uncertainty and the Brain: The Science of Neuroeconomics* (MIT Press, Cambridge, MA, 2003).
11. C. S. Sherrington, *The Integrative Action of the Nervous System* (Scribner, New York, 1906).
12. W. T. Newsome, K. H. Britten, J. A. Movshon, *Nature* **341**, 52 (1989).
13. W. T. Newsome, K. H. Britten, C. D. Salzman, J. A. Movshon, *Cold Spring Harbor Symp. Quant. Biol.* **55**, 697 (1990).
14. See also (*50*) and (*51*) for a parallel set of studies that presaged and continues to parallel many of these developments.
15. M. N. Shadlen, K. H. Britten, W. T. Newsome, J. A. Movshon, *J. Neurosci.* **16**, 1486 (1996).
16. R. A. Andersen, L. H. Snyder, D. C. Bradley, J. Xing, *Annu. Rev. Neurosci.* **20**, 303 (1997).
17. C. L. Colby, M. E. Goldberg, *Annu. Rev. Neurosci.* **22**, 319 (1999).
18. M. L. Platt, P. W. Glimcher, *Nature* **400**, 233 (1999).
19. B. Knutson, C. M. Adams, G. W. Fong, D. Hommer, *J. Neurosci.* **21**, RC159 (2001).
20. M. P. Paulus, N. Hozack, X. Zauscher, *Neuroimage* **13**, 91 (2001).
21. H. C. Breiter, I. Aharon, D. Kahneman, A. Dale, P. Shizgal, *Neuron* **30**, 619 (2001).
22. D. Kahneman, A. Tversky, *Econometrica* **47**, 263 (1979).
23. J. I. Gold, M. N. Shadlen, *Nature* **404**, 390 (2000).
24. J. I. Gold, M. N. Shadlen, *Trends Cogn. Sci.* **5**, 10 (2001).
25. M. N. Shadlen, K. H. Britten, W. T. Newsome, J. A. Movshon, *J. Neurosci.* **16**, 1486 (1996).
26. M. N. Shadlen, W. T. Newsome, *J. Neurophysiol.* **86**, 1916 (2001).
27. L. P. Sugrue, G. S. Corrado, W. T. Newsome, *Science* **304**, 1782 (2004).
28. P. Janssen, M. N. Shadlen, *Soc. Neurosci. Abstr.* **767**, 2 (2003).
29. B. Coe, K. Tomihara, M. Matsuzawa, O. Hikosaka, *J. Neurosci.* **22**, 5081 (2002).
30. J. V. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univ. Press, Princeton, NJ, 1944).
31. J. F. Nash, *Proc. Natl. Acad. Sci. U.S.A.* **36**, 48 (1950).
32. J. F. Nash, *Econometrica* **18**, 155 (1950).
33. D. J. Barraclough, M. L. Conroy, D. Lee, *Nat. Neurosci.* **7**, 404 (2004).
34. M. C. Dorris, P. W. Glimcher, *Neuron*, (in press).
35. J. F. Nash, *Ann. Math.* **54**, 286 (1951).
36. P. R. Montague *et al.*, *Neuroimage* **16**, 1159 (2002).
37. D. Ellsberg, *Q. J. Econ.* **75**, 643 (1961).
38. M. Allais, *Econometrica* **21**, 503 (1953).
39. For a review of these issues, see D. Luce, *Utility of Gains and Losses*, vol. 8 of *Scientific Psychology Series* (L. Erlbaum, Mahwah, NJ, 2000).
40. G. Gigerenzer, R. Selten, *Bounded Rationality: The Adaptive Toolbox* (MIT Press, Cambridge, MA, 2002).
41. J. Dickhaut *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3536 (2003).
42. See for example A. Bechara, H. Damasio, D. Tranel, A. Damasio, *Science* **275**, 1293 (1997).
43. J. O'Doherty, M. L. Kringelbach, E. T. Rolls, J. Hornak, C. Andrews, *Nat. Neurosci.* **4**, 95 (2001).
44. A. Rustichini, J. Dickhaut, P. Ghirardato, K. Smith, J. Pardo, in *Games and Economic Behavior*, in press.
45. W. Guth, R. Schmittberger, B. Schwarze, *J. Econ. Behav. Organ.* **3**, 347 (1982) .
46. A. Sanfey, J. K. Rilling, J. Aronson, L. E. Nystrom, J. Cohen, *Science* **300**, 1755 (2003).
47. K. McCabe, D. Houser, L. Ryan, V. Smith, T. Trouard, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11832 (2001) .
48. C. D. Frith, U. Frith, *Science* **286**, 1692 (1999).
49. H. L. Gallagher, C. D. Frith, *Trends Cogn. Sci.* **7**, 77 (2003).
50. C. R. Gallistel, *Cognition* **50**, 151 (1994).
51. K. Conover, P. Shizgal, in *Games and Economic Behavior*, in press.

arrA Is a Reliable Marker for As(V) Respiration

D. Malasarn,[1] C. W. Saltikov,[2] K. M. Campbell,[3] J. M. Santini,[4] J. G. Hering,[3] D. K. Newman[2,3]*

Bacteria play an important role in controlling the geochemistry of arsenic. A tragic example is the case of Bangladesh, where microorganisms have been implicated in the release of arsenic into drinking water supplies and the exposure of millions of people to chronic arsenic poisoning (*1*–*3*). Arsenate [As(V)] respiration [i.e., the oxidation of organic carbon, hydrogen, or sulfide coupled with As(V) reduction to arsenite] is one of the microbial processes that contributes to arsenic mobilization (*4*). It has been difficult to monitor the activity of As(V)-respiring bacteria because they are phylogenetically diverse, and this metabolic capability is not consistently present within any given clade. Here we report that a conserved functional gene, *arrA*, can be used to detect As(V)-respiratory activity in the environment.

The *arrA* gene from the Gram-negative γ-Proteobacterium, *Shewanella* species strain ANA-3, encodes for a reductase that catalyzes respiratory As(V) reduction (*5*). *arrA* is a well-conserved gene, having 61 to 100% similarity at the amino acid level when compared with seven phylogenetically diverse As(V)-respiring bacteria (supporting online text). The ArrA proteins form a unique group within the dimethyl sulfoxide (DMSO) reductase family of molybdenum-containing enzymes, which includes other terminal reductases used in microbial respiration (fig. S1).

Because of the high degree of conservation within the ArrA protein subfamily, we designed degenerate polymerase chain reaction primers, ArrAfwd (5′-AAGGTGTATGGAATAAAGCGTTTgtbgghgaytt-3′) and ArrArev (5′-CCTGTGATTTCAGGTGCCcaytyvggngt-3′), to amplify a diagnostic region of *arrA*. These primers were tested on 13 phylogenetically diverse As(V)-respiring bacteria, one As(V)-respiring archaeon, and five negative-control strains that cannot respire As(V) but that possess other genes within the DMSO reductase family (*6*). Twelve of the 13 As(V)-respiring bacteria tested positive for *arrA*; no fragments were amplified from the As(V)-respiring archaeon or negative-control strains (fig. S2). ArrAfwd and ArrArev thus appear to be reliable markers for *arrA* for the majority of As(V)-respiring bacteria.

Poorly crystalline ferric (hydr)oxide [$Fe(OH)_3$] has been shown to be the most critical sedimentary phase in controlling arsenic mobility in a variety of locales, including anaerobic sediments of the Haiwee Reservoir in Olancha, California (*7*), and Bengal delta aquifers (*8*). Accordingly, we prepared As(V)-saturated $Fe(OH)_3$ for experiments with strain ANA-3 and the mutant strain ANA-3Δ*arrA* (*6*). Both strain ANA-3 and strain ANA-3Δ*arrA* are capable of respiratory Fe(III) reduction and As(V) reduction for the purpose of detoxification using the ArsC As(V) reductase, but only ANA-3 is capable of respiratory As(V) reduction using ArrA.

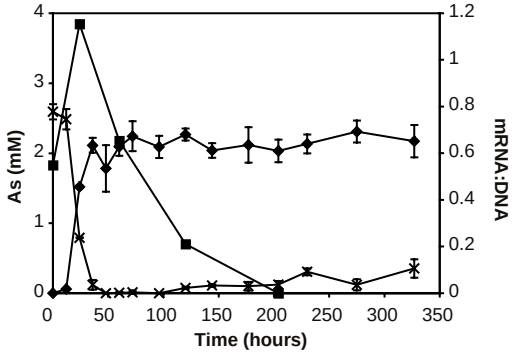


Fig. 1. *arrA* is required for As(V) reduction under iron-rich conditions. Concentrations of total As(V) (crosses) and As(III) (diamonds) are shown for samples containing *Shewanella* sp. strain ANA-3 in the presence of As(V)-saturated $Fe(OH)_3$. *arrA* expression (i.e., the mRNA:DNA ratio) is also indicated (squares). Data represent the average and standard deviation of triplicate samples, except in the case of the mRNA:DNA ratio, for which a representative data set is shown.

Reduction of As(V) occurred in samples incubated with strain ANA-3, but not in the uninoculated samples or in samples inoculated with strain ANA-3Δ*arrA*, showing that As(V) reduction is not mediated abiotically or in the absence of *arrA*, even when *arsC* is present. With strain ANA-3, the maximal expression of *arrA* (the ratio of *arrA* mRNA transcript per *arrA* gene copy number) corresponded to the fastest rate of As(V) reduction (Fig. 1). This shows that *arrA* is required to catalyze the conversion of As(V) to arsenite [As(III)] in iron-rich systems and that ArrAfwd and ArrArev can be used to track *arrA* expression.

Geochemical studies of Haiwee sediments have shown that arsenic is sorbed to $Fe(OH)_3$ and that As(III) predominates below a few centimeters (*7*). Using ArrAfwd and ArrArev, *arrA* fragments were amplified from DNA and total RNA extracted from these sediments, revealing that *arrA* is present and expressed at this site (*6*). Seven representative DNA fragments and 14 mRNA fragments were sequenced to confirm that these products were bona fide *arrA* gene fragments. These fragments ranged from ~62 to 97% identity to *arrA* sequences from *Bacillus selenitireducens*, *Chrysiogenes arsenatis*, and strain ANA-3 (fig. S3).

It is intriguing from an evolutionary perspective that *arrA* is so well conserved. On a practical level, it enables a simple molecular assay to be used to determine whether respiratory As(V) reduction is contributing to the speciation and mobilization of arsenic in a variety of environments.

References and Notes

1. C. F. Harvey *et al.*, *Science* **298**, 1602 (2002).
2. F. S. Islam *et al*, *Nature* **430**, 68 (2004).
3. A. H. Smith, E. O. Lingas, M. Rahman, *Bull. World Health Organ.* **78**, 1093 (2000).
4. R. S. Oremland, J. F. Stolz, *Science* **300**, 393 (2003).
5. C. W. Saltikov, D. K. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 10983 (2003).
6. Materials and methods are available as supporting material on *Science* Online.
7. P. E. Kneebone, P. A. O'Day, N. Jones, J. G. Hering, *Environ. Sci. Technol.* **36**, 381 (2002).
8. J. Akai *et al.*, *Appl. Geochim.* **19**, 215 (2004).
9. We thank R. Oremland, F. Rosenzweig, and C. House for sending us strains. All new *arrA* gene sequences have been deposited in GenBank. Accession numbers are available in supporting online material. Supported by the Luce and Packard Foundations (D.K.N.) and NSF award nos. BES-0201888 (J.G.H.) and DBI-0200145 (C.W.S.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5695/455/DC1
Materials and Methods
SOM Text
Figs. S1 to S3
References and Notes

6 July 2004; accepted 16 August 2004

[1]Division of Biology, [2]Division of Geological and Planetary Sciences, [3]Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125, USA. [4]Department of Microbiology, La Trobe University, 3086 Victoria, Australia.

*To whom correspondence should be addressed. E-mail: dkn@caltech.edu

Endoplasmic Reticulum Stress Links Obesity, Insulin Action, and Type 2 Diabetes

Umut Özcan,[1]* Qiong Cao,[1]* Erkan Yilmaz,[1] Ann-Hwee Lee,[2] Neal N. Iwakoshi,[2] Esra Özdelen,[1] Gürol Tuncman,[1] Cem Görgün,[1] Laurie H. Glimcher,[2,3] Gökhan S. Hotamisligil[1]†

Obesity contributes to the development of type 2 diabetes, but the underlying mechanisms are poorly understood. Using cell culture and mouse models, we show that obesity causes endoplasmic reticulum (ER) stress. This stress in turn leads to suppression of insulin receptor signaling through hyperactivation of c-Jun N-terminal kinase (JNK) and subsequent serine phosphorylation of insulin receptor substrate–1 (IRS-1). Mice deficient in X-box–binding protein–1 (XBP-1), a transcription factor that modulates the ER stress response, develop insulin resistance. These findings demonstrate that ER stress is a central feature of peripheral insulin resistance and type 2 diabetes at the molecular, cellular, and organismal levels. Pharmacologic manipulation of this pathway may offer novel opportunities for treating these common diseases.

The cluster of pathologies known as metabolic syndrome, including obesity, insulin resistance, type 2 diabetes, and cardiovascular disease, has become one of the most serious threats to human health. The dramatic increase in the incidence of obesity in most parts of the world has contributed to the emergence of this disease cluster, particularly insulin resistance and type 2 diabetes. However, understanding the molecular mechanisms underlying these individual disorders and their links with each other has been challenging.

Over the past decade, it has become clear that obesity is associated with the activation of cellular stress signaling and inflammatory pathways (*1–4*). However, the origin of this stress is not known. A key player in the cellular stress response is the ER, a membranous network that functions in the synthesis and processing of secretory and membrane proteins. Certain pathological stress conditions disrupt ER homeostasis and lead to accumulation of unfolded or misfolded proteins in the ER lumen (*5–7*). To cope with this stress, cells activate a signal transduction system linking the ER lumen with the cytoplasm and nucleus, called the unfolded protein response (UPR) (*5–7*). Among the conditions that trigger ER stress are glucose or nutrient deprivation, viral infections, lipids, increased synthesis of secretory proteins, and expression of mutant or misfolded proteins (*8–10*).

Several of these conditions occur in obesity. Specifically, obesity increases the demand on the synthetic machinery of the cells in many secretory organ systems. Obesity is also associated with mechanical stress, excess lipid accumulation, abnormalities in intracellular energy fluxes, and nutrient availability. In light of these observations, we postulated that obesity may be a chronic stimulus for ER stress in peripheral tissues and that perhaps ER stress is a core mechanism involved in triggering insulin resistance and type 2 diabetes.

Induction of ER stress in obesity. To examine whether ER stress is increased in obesity, we investigated the expression patterns of several molecular indicators of ER stress in dietary [high-fat diet (HFD)–induced] and genetic (*ob/ob*) models of murine obesity. The pancreatic ER kinase or PKR-like kinase (PERK) is an ER transmembrane protein kinase that phosphorylates the α subunit of translation initiation factor 2 (eIF2α) in response to ER stress. The phosphorylation status of PERK and eIF2α is therefore a key indicator of the presence of ER stress (*11–13*). We determined the phosphorylation status of PERK (Thr^{980}) and eIF2α (Ser^{51}) using phospho-specific antibodies. These experiments demonstrated increased PERK and eIF2α phosphorylation in liver extracts of obese mice compared with lean controls (Fig. 1, A and B). The activity of c-Jun N-terminal kinase (JNK) is also increased by ER stress (*14*). Consistent with earlier observations (*3*), total JNK activity, indicated by c-Jun phosphorylation, was also dramatically elevated in the obese mice (Fig. 1, A and B).

The 78-kD glucose-regulated/binding immunoglobulin protein (GRP78) is an ER chaperone whose expression is increased upon ER stress (*7*). The GRP78 mRNA levels were elevated in the liver tissue of obese mice compared with matched lean controls (Fig. 1, C and D). Because GRP78 expression is responsive to glucose (*15*), we tested whether this up-regulation might simply be due to increasing glucose levels. Treatment of cultured rat Fao liver cells with high levels of glucose resulted in reduced GRP78 expression (fig. S1A). Similarly, GRP78 levels were not increased in a mouse model of hyperglycemia (fig. S1B), which indicates that regulation in obesity is unlikely to be related to glycemia alone.

We also tested adipose and muscle tissues, important sites for metabolic ho-

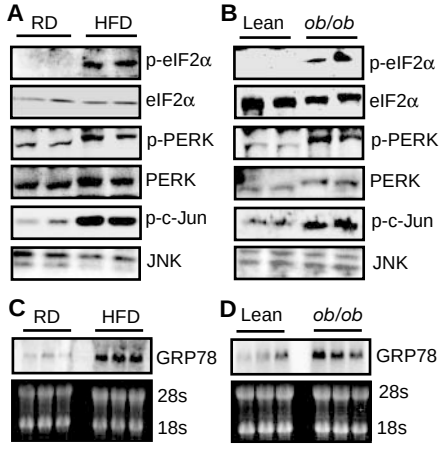


Fig. 1. Increased ER stress in obesity. Dietary (HFD-induced) and genetic (*ob/ob*) models of mouse obesity were used to examine markers of ER stress in liver tissue compared with age- and sex-matched lean controls. (**A**) ER stress markers including eIF2α phosphorylation (p-eIF2α), PERK phosphorylation (p-PERK), and JNK activity (p-c-Jun) were examined in the liver samples of the male mice (C57BL/6) that were kept either on regular diet (RD) or high-fat diet (HFD) for 16 weeks. (**B**) Examination of the same ER stress markers in the livers of male *ob/ob* and wild-type (WT) lean mice at the age of 12 to 14 weeks. (**C**) Northern blot analysis of GRP78 mRNA in the livers of mice with dietary-induced obesity and lean controls. (**D**) Northern blot analysis of GRP78 mRNA in the livers of *ob/ob* and WT lean mice. Ethidium bromide staining is shown as a control for loading and integrity of RNA.

[1]Department of Genetics and Complex Diseases, [2]Department of Immunology and Infectious Diseases, Harvard School of Public Health, [3]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: ghotamis@hsph.harvard.edu

meostasis, for indications of ER stress in obesity. As in liver, PERK phosphorylation, JNK activity, and GRP78 expression were all significantly increased in adipose tissue of obese animals compared with lean controls (fig. S2, A to C). However, no indication for ER stress was evident in the muscle tissue of obese animals (*16*). Taken together, these results indicate that obesity is associated with induction of ER stress predominantly in liver and adipose tissues.

ER stress inhibits insulin action in liver cells. To investigate whether ER stress interferes with insulin action, we pretreated Fao liver cells with tunicamycin or thapsigargin, agents commonly used to induce ER stress. Tunicamycin significantly decreased insulin-stimulated tyrosine phosphorylation of insulin receptor substrate 1 (IRS-1) (Fig. 2, A and B), and it also produced an increase in the molecular weight of IRS-1 (Fig. 2A). IRS-1 is a substrate for insulin receptor tyrosine kinase, and serine phosphorylation of IRS-1, particularly mediated by JNK, reduces insulin receptor signaling (*3*). Pretreatment of Fao cells with tunicamycin produced a significant increase in serine phosphorylation of IRS-1 (Fig. 2, A and B). Tunicamycin pretreatment also suppressed insulin-induced Akt phosphorylation, a more distal event in the insulin receptor signaling pathway (Fig. 2, A and B). Similar results were also obtained after treatment with thapsigargin (fig. S3A), which was independent of alterations in cellular calcium levels (fig. S3B). Hence, experimental ER stress inhibits insulin action.

We next examined the role of JNK in ER stress–induced IRS-1 serine phosphorylation and inhibition of insulin-stimulated IRS-1 tyrosine phosphorylation. Inhibition of JNK activity with the synthetic inhibitor, SP600125 (*17*), reversed the ER stress–induced serine phosphorylation of IRS-1 (Fig. 2, C and D). Pretreatment of Fao cells with a highly specific inhibitory peptide derived from the JNK-binding protein, JIP (*18*), also completely preserved insulin receptor signaling in cells exposed to tunicamycin (Fig. 2, E and F). Similar results were obtained with the synthetic JNK inhibitor, SP600125 (*16*). These results indicate that ER stress promotes a JNK-dependent serine phosphorylation of IRS-1, which in turn inhibits insulin receptor signaling.

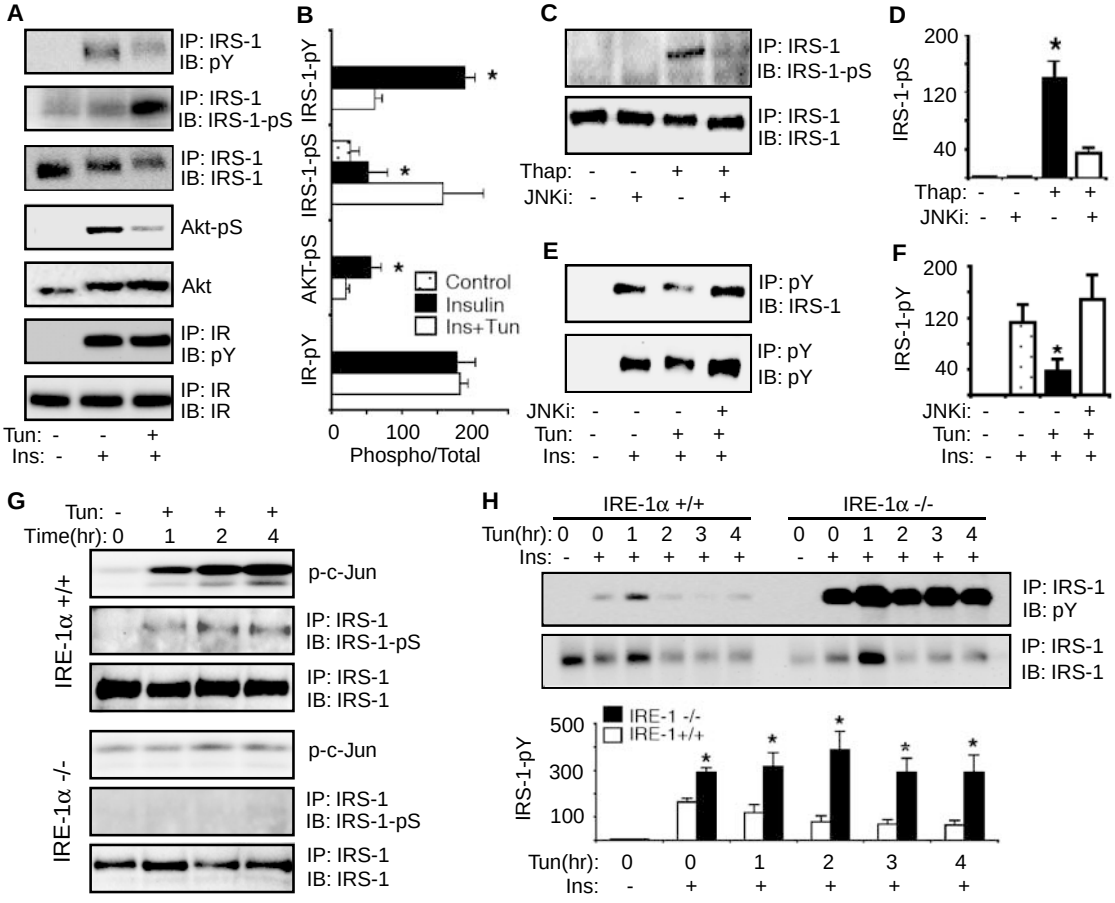


Fig. 2. Induction of ER stress impairs insulin action through JNK-mediated phosphorylation of IRS-1. (**A**) ER stress was induced in Fao liver cells by a 3-hour treatment with 5 μg/ml tunicamycin (Tun). Cells were subsequently stimulated with insulin (Ins). IRS-1 tyrosine (pY) and Ser^{307} (pS) phosphorylation, Akt Ser^{473} (Akt-pS) phosphorylation, insulin receptor (IR) tyrosine phosphorylation, and their total protein levels were examined either with immunoprecipitation (IP) followed by immunoblotting (IB) or by direct immunoblotting. (**B**) Quantification of IRS-1 (tyrosine and Ser^{307}), Akt (Ser^{473}), and IR (tyrosine) phosphorylation under the experimental conditions described in (A) with normalization to protein levels for each molecule. (**C**) Inhibition of ER stress–induced (300 nM thapsigargin for 4 hours) Ser^{307} phosphorylation of IRS-1 by JNK-1 inhibitor, SP600125 (JNKi, 25 μM). (**D**) Quantification of IRS-1 Ser^{307} phosphorylation under the conditions described in (C). (**E**) Reversal of ER stress–induced inhibition of insulin-stimulated tyrosine phosphorylation (pY) of IRS-1 by a peptide JNK inhibitor. (**F**) Quantification of insulin-induced IRS-1 tyrosine phosphorylation levels described in (E). (**G**) JNK activity (p-c-Jun), Ser^{307} phosphorylation of IRS-1, and total IRS-1 levels at indicated times after tunicamycin treatment (Tun, 10 μg/ml) in IRE-1α$^{+/+}$ and IRE-1α$^{-/-}$ fibroblasts. (**H**) Insulin-stimulated IRS-1 tyrosine phosphorylation and total IRS-1 levels after tunicamycin treatment (Tun, 10 μg/ml) in IRE-1α$^{+/+}$ and IRE-1α$^{-/-}$ fibroblasts. Quantification of insulin-induced IRS-1 tyrosine phosphorylation levels in IRE-1α$^{+/+}$ and IRE-1α$^{-/-}$ cells is displayed in the bottom of the panel. All graphs show means ± SEM from at least two independent experiments, and statistical significance ($P < 0.005$) from the controls is indicated by an asterisk (*).

Inositol-requiring kinase–1α (IRE-1α) plays a crucial role in insulin receptor signaling. In the presence of ER stress, increased phosphorylation of IRE-1α leads to recruitment of tumor necrosis factor receptor–associated factor 2 (TRAF2) protein and activation of JNK (*14*). To address whether ER stress–induced insulin resistance is dependent on intact IRE-1α, we measured JNK activation, IRS-1 serine phosphorylation, and insulin receptor signaling after exposure of IRE-1α$^{-/-}$ and wild-type fibroblasts to tunicamycin. In the wild-type, but not IRE-1α$^{-/-}$ cells, induction of ER stress by tunicamycin resulted in strong activation of JNK (Fig. 2G). Tunicamycin also stimulated phosphorylation of IRS-1 at the Ser^{307} residue in wild-type, but not IRE-1α$^{-/-}$, fibroblasts (Fig. 2G). It is noteworthy that tunicamycin inhibited insulin-stimulated tyrosine phosphorylation of IRS-1 in the wild-type cells, whereas no such effect was detected in the IRE-1α$^{-/-}$ cells (Fig. 2H). The level of insulin-induced tyrosine phosphorylation of IRS-1 was dramatically higher in IRE-1α$^{-/-}$ cells, despite lower total IRS-1 protein levels (Fig. 2H). These results demonstrate that ER stress–induced inhibition of insulin action is mediated by an IRE-1α– and JNK–dependent protein kinase cascade.

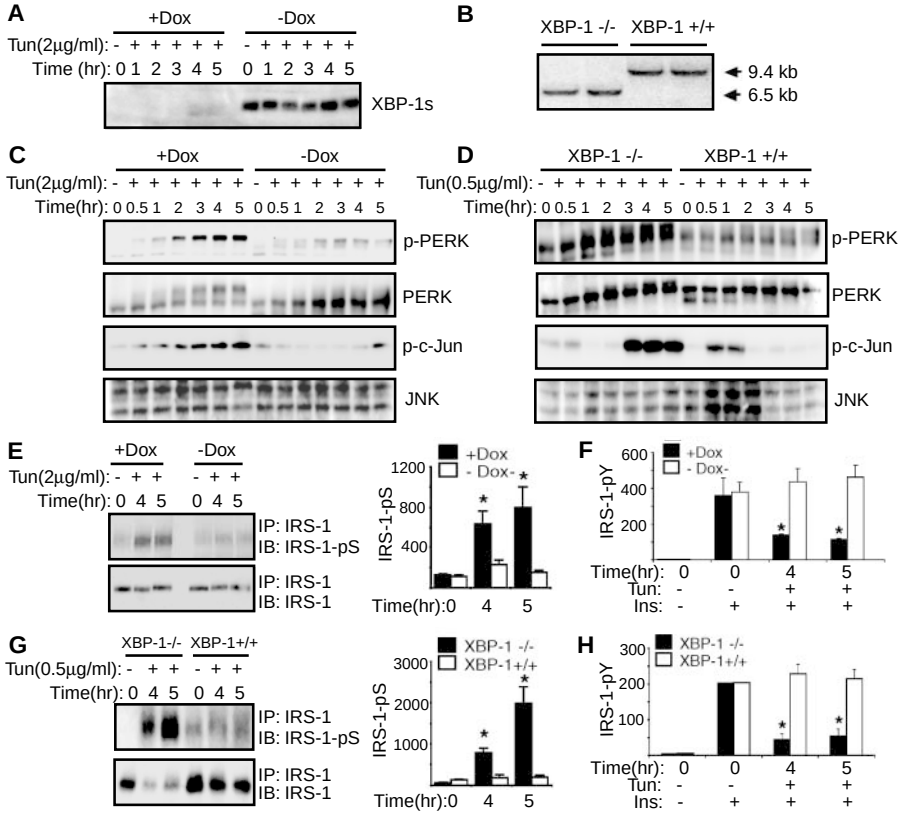


Fig. 3. Alteration of the ER stress response by manipulation of XBP-1 levels modulates insulin receptor signaling. ER stress responses in cells overexpressing XBP-1s, XBP-1$^{-/-}$ cells, and their controls. (**A**) Induction of exogenous XBP-1s expression on removal of doxycycline in MEFs. (**B**) Southern blot analysis of XBP-1$^{-/-}$ MEFs and their WT controls (9.4 kb) and targeted (6.5 kb) alleles. (**C**) PERK phosphorylation (p-PERK) and JNK activity (p-c-Jun) in cells overexpressing XBP-1s and control cells (–Dox and +Dox, respectively) after tunicamycin treatment (Tun, 2 μg/ml). (**D**) PERK phosphorylation and JNK activity after low-dose tunicamycin treatment (Tun, 0.5 μg/ml) in XBP-1$^{-/-}$ MEFs and their WT controls. (**E**) IRS-1 Ser^{307} phosphorylation (pS) after tunicamycin treatment (Tun, 2 μg/ml) in cells overexpressing XBP-1s and control cells (–Dox and +Dox, respectively), detected by using immunoprecipitation (IP) of IRS-1 followed by immunoblotting (IB) with an IRS-1 phosphoserine 307–specific antibody. The graph next to the blots shows the quantification of IRS-1 Ser^{307} phosphorylation under the conditions described in (E). (**F**) Insulin-stimulated tyrosine phosphorylation of IRS-1 in cells overexpressing XBP-1s and control cells, with or without tunicamycin treatment (Tun, 2 μg/ml). The ratio of IRS-1 tyrosine phosphorylation to total IRS-1 level was summarized from independent experiments and presented in the graph. (**G**) IRS-1 Ser^{307} phosphorylation after tunicamycin treatment (Tun, 0.5 μg/ml) in XBP-1$^{-/-}$ cells and WT controls was detected as described in (C). The graph next to the blots shows the quantification of IRS-1 Ser^{307} phosphorylation under conditions described in (G). (**H**) Insulin-stimulated tyrosine phosphorylation of IRS-1 in XBP-1$^{-/-}$ and WT control cells with or without tunicamycin treatment (Tun, 0.5 μg/ml). The ratio of IRS-1 tyrosine phosphorylation to total IRS-1 level was summarized from independent experiments and presented in the graph. All graphs show means ± SEM from at least two independent experiments, and statistical significance from the controls is indicated by * with $P < 0.005$.

Manipulation of X-box–binding protein–1 (XBP-1) levels alters insulin receptor signaling. The transcription factor XBP-1 is a bZIP protein. The spliced or processed form of XBP-1 (XBP-1s) is a key factor in ER stress through transcriptional regulation of an array of genes, including molecular chaperones (*19–22*). We therefore reasoned that modulation of XBP-1s levels in cells should alter insulin action via its potential impact on the magnitude of the ER stress responses. To test this possibility, we established XBP-1 gain- and loss-of-function cellular models. First, we established an inducible gene expression system where exogenous XBP-1s is expressed only in the absence of tetracycline/doxycycline (Fig. 3A). In parallel, we also studied mouse embryo fibroblasts (MEFs) derived from XBP-1$^{-/-}$ mice (Fig. 3B). In fibroblasts without exogenous XBP-1s expression, tunicamycin treatment (2 μg/ml) resulted in PERK phosphorylation starting at 30 min and peaking at 3 to 4 hours, associated with a mobility shift characteristic of PERK phosphorylation (Fig. 3C). In these cells, there was also a rapid and robust activation of JNK in response to ER stress (Fig. 3C). When XBP-1s expression was induced, there was a dramatic reduction in both PERK phosphorylation and JNK activation after tunicamycin treatment (Fig. 3C). Hence, overexpression of XBP-1s rendered wild-type cells refractory to ER stress. Similar experiments performed in XBP-1$^{-/-}$ MEFs revealed an opposite pattern (Fig. 3D). XBP-1$^{-/-}$ MEFs mounted strong ER stress responses even when treated with a low dose of tunicamycin (0.5 μg/ml), which failed to stimulate significant ER stress in wild-type cells (Fig. 3D). Under these conditions, PERK phosphorylation and JNK activation levels in XBP-1$^{-/-}$ MEFs were significantly higher than those seen in wild-type controls (Fig. 3D), which indicates that XBP-1$^{-/-}$ cells are prone to ER stress. Thus, alterations in the levels of cellular XBP-1s protein result in alterations in the ER stress responses.

Next, we examined whether these differences in the ER stress responses produced alterations in insulin action as assessed by IRS-1 serine phosphorylation and insulin-stimulated IRS-1 tyrosine phosphorylation. Tunicamycin-induced IRS-1 serine phosphorylation was significantly reduced in fibroblasts exogenously expressing XBP-1s, compared with that of control cells (Fig. 3E). On insulin stimulation, the extent of IRS-1 tyrosine phosphorylation was significantly

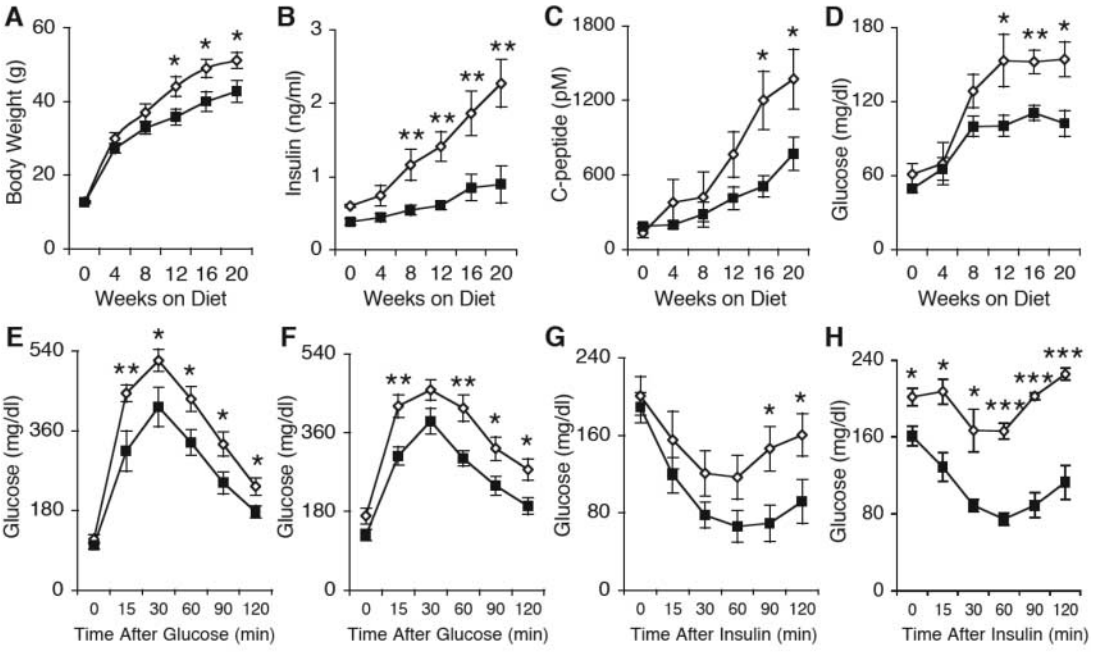


Fig. 4. Glucose homeostasis in XBP-$1^{+/-}$ mice fed HFD. The XBP-$1^{+/-}$ (◇) and XBP-$1^{+/+}$ (■) mice were fed HFD immediately after weaning at 3 weeks of age. Total body weight (**A**), fasting blood insulin (**B**), C-peptide (**C**), and glucose (**D**) levels were measured in the XBP-$1^{+/-}$ and XBP-$1^{+/+}$ mice during the course of HFD. GTT were performed after 7 (**E**) and 16 (**F**) weeks of HFD in XBP-$1^{+/-}$ and XBP-$1^{+/+}$ mice. ITT were performed after 8 (**G**) and 17 (**H**) weeks of HFD in XBP-$1^{+/-}$ (n = 11) and XBP-$1^{+/+}$ (n = 8) mice. Data are shown as means ± SEM. Statistical significance in two-tailed Student's t test is indicated by $^{*}P \leq 0.05$, $^{**}P \leq 0.005$, and $^{***}P \leq 0.0005$. XBP-$1^{+/-}$ and XBP-$1^{+/+}$ groups are also compared by ANOVA (A to H).

higher in cells overexpressing XBP-1s, compared with controls (Fig. 3F). In contrast, IRS-1 serine phosphorylation was strongly induced in XBP-$1^{-/-}$ MEFs compared with XBP-$1^{+/+}$ controls even at low doses of tunicamycin treatment (0.5 μg/ml) (Fig. 3G). After insulin stimulation, the amount of IRS-1 tyrosine phosphorylation was significantly decreased in tunicamycin-treated XBP-$1^{-/-}$ cells compared with tunicamycin-treated wild-type controls (Fig. 3H). Insulin-stimulated tyrosine phosphorylation of the insulin receptor was normal in these cells (fig. S4).

XBP-$1^{+/-}$ mice show impaired glucose homeostasis. Complete XBP-1 deficiency results in embryonic lethality (*23*). To investigate the role of XBP-1 in ER stress, insulin sensitivity, and systemic glucose metabolism in vivo, we studied BALB/c-XBP-$1^{+/-}$ mice with a null mutation in one XBP-1 allele. We chose mice on the BALB/c genetic background, because this strain exhibits strong resistance to obesity-induced alterations in systemic glucose metabolism. Based on our results with cellular systems, we hypothesized that XBP-1 deficiency would predispose mice to the development of insulin resistance and type 2 diabetes.

We fed XBP-$1^{+/-}$ mice and their wild-type littermates a HFD at 3 weeks of age. In parallel, control mice of both genotypes were placed on laboratory feed, a regular diet. The total body weights of both genotypes were similar with regular diet and until 12 weeks of age when fed HFD. After this period, the XBP-$1^{+/-}$animals fed HFD exhibited a small, but significant, increase in body weight (Fig. 4A). Serum levels of leptin, adiponectin, and triglycerides did not exhibit any statistically significant differences between the genotypes measured after 16 weeks of HFD (fig. S5).

Fed HFD, XBP-$1^{+/-}$ mice developed continuous and progressive hyperinsulinemia evident as early as 4 weeks (Fig. 4B). Insulin levels continued to increase in XBP-$1^{+/-}$ mice for the duration of the experiment. Blood insulin levels in XBP-$1^{+/+}$ mice were significantly lower than those in XBP-$1^{+/-}$ littermates (Fig. 4B). C-peptide levels were also significantly higher in XBP-$1^{+/-}$ animals than in wild-type controls (Fig. 4C). Blood glucose levels also began to rise in the XBP-$1^{+/-}$ mice fed HFD starting at 8 weeks and remained high until the conclusion of the experiment at 20 weeks (Fig. 4D). This pattern was the same in both fasted (Fig. 4D) and fed (*16*) states. The rise in blood glucose in the face of hyperinsulinemia in the mice fed HFD is a strong indicator of the development of peripheral insulin resistance.

To investigate systemic insulin sensitivity, we performed glucose tolerance tests (GTT) and insulin tolerance tests (ITT) in XBP-$1^{+/-}$ mice and XBP-$1^{+/+}$ controls. Exposure to HFD resulted in significant glucose intolerance in XBP-$1^{+/-}$ mice. After 7 weeks of HFD, XBP-$1^{+/-}$ mice showed significantly higher glucose levels on glucose challenge than XBP-$1^{+/+}$ mice (Fig. 4E). This glucose intolerance continued to be evident in XBP-$1^{+/-}$ mice compared with wild-type mice after 16 weeks of HFD (Fig. 4F). During ITT, the hypoglycemic response to insulin was also significantly lower in XBP-$1^{+/-}$ mice compared with XBP-$1^{+/+}$ littermates at 8 weeks of HFD (Fig. 4G), and this reduced responsiveness continued to be evident after 17 weeks of HFD (Fig. 4H). Examination of islets morphology and function did not reveal significant differences between genotypes (fig. S6). Hence, loss of an XBP-1 allele predisposes mice to diet-induced peripheral insulin resistance and type 2 diabetes.

Increased ER stress and impaired insulin signaling in XBP-$1^{+/-}$ mice. Our experiments with cultured cells demonstrated an increase in ER stress and a decrease in insulin signaling capacity in XBP-1–deficient cells, as well as reversal of these phenotypes on expression of high levels of XBP-1s. If this mechanism is the basis of the insulin resistance seen in XBP-$1^{+/-}$ mice, these animals should exhibit high levels of ER stress coupled with impaired insulin receptor signaling. To test this, we first examined PERK phosphorylation and JNK activity in the livers of obese XBP-$1^{+/-}$ and wild-type mice. These experiments revealed an increase in PERK levels and seemingly an increase in liver PERK phosphorylation in obese XBP-$1^{+/-}$ mice compared with wild-type controls fed HFD (Fig. 5A). There was a significant increase in JNK activity in XBP-$1^{+/-}$ mice compared with wild-type controls (Fig. 5B). Consistent with these results, Ser^{307} phosphorylation of IRS-1 was also increased in XBP-$1^{+/-}$ mice compared with wild-type controls fed HFD (Fig. 5C). Finally, we studied in vivo insulin-stimulated, insulin receptor–signaling capacity in these mice.

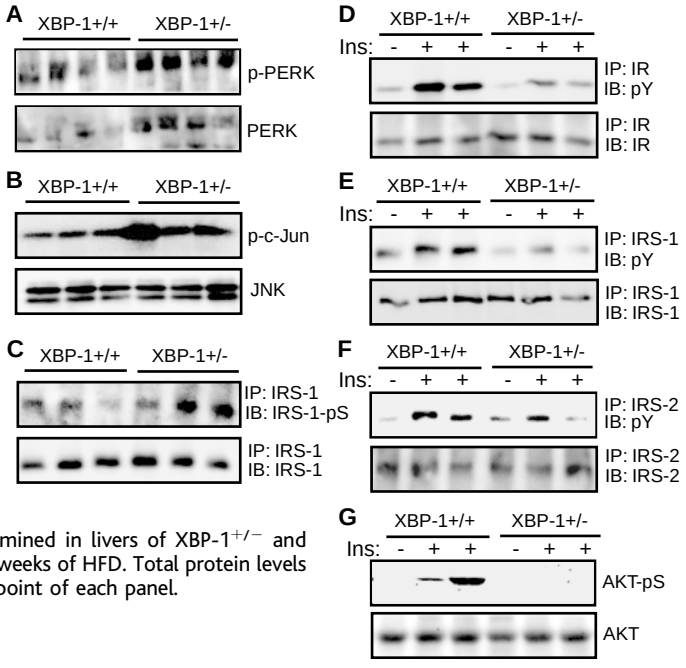


Fig. 5. ER stress and insulin receptor signaling in XBP-$1^{+/-}$ mice. PERK phosphorylation (p-PERK) (**A**), JNK activity (p-c-Jun) (**B**), and Ser^{307} phosphorylation (pS) of IRS-1 (**C**) were examined in the livers of XBP-$1^{+/-}$ and XBP-$1^{+/+}$ mice after 16 weeks of HFD. After infusion of insulin (1 U/kg) through the portal vein, insulin receptor (IR) tyrosine phosphorylation (pY) (**D**), IRS-1 tyrosine phosphorylation (**E**), IRS-2 tyrosine phosphorylation (**F**), and Akt Ser^{473} phosphorylation (pS) (**G**) were examined in livers of XBP-$1^{+/-}$ and XBP-$1^{+/+}$ mice after 16 weeks of HFD. Total protein levels are shown in the lower point of each panel.

There was no detectable difference in any of the insulin receptor–signaling components in liver and adipose tissues between genotypes taking regular diet (fig. S7). However, after exposure to HFD, major components of insulin receptor signaling in the liver, including insulin-stimulated insulin receptor, IRS-1, and IRS-2 tyrosine- and Akt serine-phosphorylation, were all decreased in XBP-$1^{+/-}$ mice compared with wild-type controls (Fig. 5, D to G). A similar suppression of insulin receptor signaling was also evident in the adipose tissues of XBP-$1^{+/-}$ mice compared with XBP-$1^{+/+}$ mice fed HFD (fig. S8). The suppression of IR tyrosine phosphorylation in XBP-$1^{+/-}$ mice differs from the observations made in XBP-$1^{-/-}$ cells, where ER stress inhibited insulin action after the receptor signal in the pathway. It is likely that this difference reflects the effects of chronic hyperinsulinemia in vivo on insulin receptors. Hence, our data demonstrate the link between ER stress and insulin action in vivo but are not conclusive in determining the exact locus in insulin receptor signaling pathway that is targeted through this mechanism.

Discussion. In this study, we identify ER stress as a molecular link between obesity, the deterioration of insulin action, and the development of type 2 diabetes. Induction of ER stress or reduction in the compensatory capacity through down-regulation of XBP-1 leads to suppression of insulin receptor signaling in intact cells via IRE-1α–dependent activation of JNK. Experiments with mouse models also yielded data consistent with the link between ER stress and systemic insulin action. Deletion of an XBP-1 allele in mice leads to enhanced ER stress, hyperactivation of JNK, reduced insulin receptor signaling, systemic insulin resistance, and type 2 diabetes.

Our findings point to a fundamental mechanism underlying the molecular sensing of obesity-induced metabolic stress by the ER and inhibition of insulin action that ultimately leads to insulin resistance and type 2 diabetes. We therefore postulate that ER stress underlies the emergence of the stress and inflammatory responses in obesity and the integrated deterioration of systemic glucose homeostasis.

Although our results in this study predominantly point to a role for ER stress in peripheral insulin resistance, earlier studies have linked ER stress with islet function and survival. For example, PERK$^{-/-}$ mice exhibit a phenotype resembling type 1 diabetes resulting from pancreatic islet destruction soon after birth (*24*). PERK mutations also cause a rare inherited form of type 1 diabetes in humans (*25*). Loss of eIF2α phosphorylation by targeted mutation of serine 51 residue of eIF2α to alanine also leads to alterations in pancreatic beta cell function, in addition to its impact on liver gluconeogenesis (*11*, *26*). Therefore, we propose that the effect of chronic ER stress on glucose homeostasis in obesity could represent a central and integrating mechanism underlying both peripheral insulin resistance and impaired insulin secretion.

The critical role of ER stress responses in insulin action may represent a mechanism conserved by evolution, whereby stress signals are integrated with metabolic regulatory pathways through the ER. This integration could have been advantageous, because proper regulation of energy fluxes and the suppression of major anabolic pathways might have been favorable during acute stress, pathogen invasion, and immune responses. Hence, the trait would propagate through natural selection. However, in the presence of chronic ER stress, such as we see in obesity, the effect of ER stress on metabolic regulation would lead to the development of insulin resistance and, eventually, type 2 diabetes. In terms of therapeutics, our findings suggest that interventions that regulate the ER stress response offer new opportunities for preventing and treating type 2 diabetes.

References and Notes

1. G. S. Hotamisligil, in *Diabetes Mellitus*, D. LeRoith, S. I. Taylor, J. M. Olefsky, Eds. (Lippincott Williams & Wilkins, Philadelphia, 2003), pp. 953–962.
2. K. T. Uysal, S. M. Wiesbrock, M. W. Marino, G. S. Hotamisligil, *Nature* **389**, 610 (1997).
3. J. Hirosumi *et al.*, *Nature* **420**, 333 (2002).
4. M. Yuan *et al.*, *Science* **293**, 1673 (2001).
5. R. Y. Hampton, *Curr. Biol.* **10**, R518 (2000).
6. K. Mori, *Cell* **101**, 451 (2000).
7. H. P. Harding, M. Calfon, F. Urano, I. Novoa, D. Ron, *Annu. Rev. Cell Dev. Biol.* **18**, 575 (2002).
8. Y. Ma, L. M. Hendershot, *Cell* **107**, 827 (2001).
9. R. J. Kaufman *et al.*, *Nature Rev. Mol. Cell Biol.* **3**, 411 (2002).
10. I. Kharroubi *et al.*, *Endocrinology* (2004); published online 5 August 2004 (10.1210/en.2004-0478).
11. Y. Shi, S. I. Taylor, S. L. Tan, N. Sonenberg, *Endocr. Rev.* **24**, 91 (2003).
12. Y. Shi *et al.*, *Mol. Cell. Biol.* **18**, 7499 (1998).
13. H. P. Harding, Y. Zhang, D. Ron, *Nature* **397**, 271 (1999).
14. F. Urano *et al.*, *Science* **287**, 664 (2000).
15. R. P. Shiu, J. Pouyssegur, I. Pastan, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 3840 (1977).
16. U. Özcan, G. S. Hotamisligil, unpublished observations.
17. B. L. Bennett *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13681 (2001).
18. R. K. Barr, T. S. Kendrick, M. A. Bogoyevitch, *J. Biol. Chem.* **277**, 10987 (2002).
19. M. Calfon *et al.*, *Nature* **415**, 92 (2002).
20. X. Shen *et al.*, *Cell* **107**, 893 (2001).
21. H. Yoshida, T. Matsui, A. Yamamoto, T. Okada, K. Mori, *Cell* **107**, 881 (2001).
22. A. H. Lee, N. N. Iwakoshi, L. H. Glimcher, *Mol. Cell. Biol.* **23**, 7448 (2003).
23. A. M. Reimold *et al.*, *Genes Dev.* **14**, 152 (2000).
24. H. P. Harding *et al.*, *Mol. Cell* **7**, 1153 (2001).
25. M. Delepine *et al.*, *Nature Genet.* **25**, 406 (2000).
26. D. Scheuner *et al.*, *Mol. Cell* **7**, 1165 (2001).

27. We thank the Hotamisligil laboratory for their contributions and J. Gound and L. Beppu for technical assistance. Supported in part by NIH grants AI32412 (L.H.G.), DK52539 (G.S.H.), American Diabetes Association (G.S.H.), PO5-CA100707 (L.H.G., A.H.L.), an Irvington Institute Postdoctoral Fellowship Award (N.I.), an NIH training grant T32-DK07703 (Q.C.), and a postdoctoral fellowship from the Iaccoca Foundation (G.T.). L.H.G. holds equity in MannKind Corporation, which has licensed the XBP-1 technology.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5695/457/DC1
Materials and Methods
Figs. S1 to S8
References

23 July 2004; accepted 9 September 2004

A Bayesian Truth Serum for Subjective Data

Dražen Prelec

Subjective judgments, an essential information source for science and policy, are problematic because there are no public criteria for assessing judgmental truthfulness. I present a scoring method for eliciting truthful subjective data in situations where objective truth is unknowable. The method assigns high scores not to the most common answers but to the answers that are more common than collectively predicted, with predictions drawn from the same population. This simple adjustment in the scoring criterion removes all bias in favor of consensus: Truthful answers maximize expected score even for respondents who believe that their answer represents a minority view.

Subjective judgment from expert and lay sources is woven into all human knowledge. Surveys of behaviors, attitudes, and intentions are a research staple in political science, psychology, sociology, and economics (*1*). Subjective expert judgment drives environmental risk analysis, business forecasts, historical inferences, and artistic and legal interpretations (*2*).

The value of subjective data is limited by its quality at the source—the thought process of an individual respondent or expert. Quality would plausibly be enhanced if respondents felt as if their answers were being evaluated by an omniscient scorer who knew the truth (*3*). This is the situation with tests of objective knowledge, where success is defined as agreement with the scorer's answer key, or in the case of forecasts, an observable outcome (*4*). Such evaluations are rarely appropriate in social science, because the scientist is reluctant to impose a particular definition of truth, even if one were available (*5*).

Here, I present a method of eliciting subjective information, designed for situations where objective truth is intrinsically or practically unknowable (*6*). The method consists of an "information-scoring" system that induces truthful answers from a sample of rational (i.e., Bayesian) expected value–maximizing respondents. Unlike other Bayesian elicitation mechanisms (*7*–*9*), the method does not assume that the researcher knows the probabilistic relationship between different responses. Hence, it can be applied to previously unasked questions, by a researcher who is a complete outsider for the domain. Unlike earlier approaches to "test theory without an answer key" (*5*), or the Delphi method (*10*), it does not privilege the consensus answer. Hence, there is no reason for respondents to bias their answer toward the likely group mean. Truthful responding remains the correct strategy even for someone who is sure that their answer represents a minority view.

Instead of using consensus as a truth criterion, my method assigns high scores to answers that are more common than collectively predicted, with predictions drawn from the same population that generates the answers. Such responses are "surprisingly common," and the associated numerical index is called an information score. This adjustment in the target criterion removes the bias inherent in consensus-based methods and levels the playing field between typical and unusual opinions.

The scoring works at the level of a single question. For example, we might ask: (i) What is your probability estimate that humanity will survive past the year 2100 (100-point probability scale)? (ii) Will you vote in the next presidential election (Definitely/Probably/Probably Not/Definitely Not)? (iii) Have you had more than 20 sexual partners over the past year (Yes/No)? (iv) Is Picasso your favorite 20th-century painter (Yes/No)?

Each respondent provides a personal answer and also a prediction of the empirical distribution of answers (i.e., the fraction of people endorsing each answer). Predictions are scored for accuracy, that is, for how well they match the empirical frequencies. The personal answers, which are the main object of interest, are scored for being surprisingly common. An answer endorsed by 10% of the population against a predicted frequency of 5% would be surprisingly common and would receive a high information score; if predictions averaged 25%, it would be a surprisingly uncommon answer, and hence receive a low score.

The surprisingly common criterion exploits an overlooked implication of Bayesian reasoning about population frequencies. Namely, in most situations, one should expect that others will underestimate the true frequency of one's own opinion or personal characteristic. This implication is a corollary to the more usual Bayesian argument that the highest predictions of the frequency of a given opinion or characteristic in the population should come from individuals who hold that opinion or characteristic, because holding the opinion constitutes a valid and favorable signal about its general popularity (*11*, *12*). People who, for example, rate Picasso as their favorite should—and usually do (*13*)—give higher estimates of the percentage of the population who shares that opinion, because their own feelings are an informative "sample of one" (*14*). It follows, then, that Picasso lovers, who have reason to believe that their best estimate of Picasso popularity is high compared with others' estimates, should conclude that the true popularity of Picasso is underestimated by the population. Hence, one's true opinion is also the opinion that has the best chance of being surprisingly common.

The validity of this conclusion does not depend on whether the personally truthful answer is believed to be rare or widely shared. For example, a male who has had more than 20 sexual partners [answering question (iii)] may feel that few people fall in this promiscuous category. Nevertheless, according to Bayesian reasoning, he should expect that his personal estimate of the percentage (e.g., 5%) will be somewhat higher than the average of estimates collected from the population as a whole (e.g., 2%). The fact that he has had more than 20 sexual partners is evidence that the general population, which includes persons with fewer partners, will underestimate the prevalence of this profile.

Truth-telling is individually rational in the sense that a truthful answer maximizes expected information score, assuming that everyone is responding truthfully [hence, it is a Bayesian Nash equilibrium (*15*)]. It is also collectively rational in the sense that no other equilibrium provides a higher expected information score, for any respondent. In actual applications of the method, one would not teach respondents the mathematics of scoring or explain the notion of equilibrium. Rather, one would like to be able to tell them that truthful answers will maximize their expected scores, and that in arriving at their personal true answer they are free to ignore what other respondents might say. The equilibrium analysis confirms that under cer-

Massachusetts Institute of Technology, Sloan School of Management, E56-320, 38 Memorial Drive, Cambridge, MA 02139, USA. E-mail: dprelec@mit.edu

tain conditions one can make such a claim honestly.

The equilibrium results rest on two assumptions. First, the sample of respondents must be sufficiently large so that a single answer cannot appreciably affect empirical frequencies (*16*). The results do hold for large finite populations but are simpler to state for a countably infinite population, as is done here. Respondents are indexed by $r \in \{1,2,\ldots\}$, and their truthful answer to a m multiple-choice question by $t^r = (t^r_1,..,t^r_m)$ ($t^r_k \in \{0,1\}$, $\Sigma_k x^r_k = 1$). t^r_k is thus an indicator variable that has a value of one or zero depending on whether answer k is or is not the truthful answer of respondent r. The truthful answer is also called a personal opinion or characteristic.

Second, respondents treat personal opinions as an "impersonally informative" signal about the population distribution, which is an unknown parameter, $\omega = (\omega_1,..,\omega_m) \in \Omega$ (*17*). Formally, I assume common knowledge (*18*) by respondents that all posterior beliefs, $p(\omega|t^r)$, are consistent with Bayesian updating from a single distribution over ω, also called a common prior, $p(\omega)$, and that: $p(\omega|t^r) = p(\omega|t^s)$ if and only if $t^r = t^s$. Opinions thus provide evidence about ω, but the inference is impersonal: Respondents believe that others sharing their opinion will draw the same inference about population frequencies (*19*). One can therefore denote a generic respondent with opinion j by t_j and suppress the respondent superscript from joint and conditional probabilities: $Prob\{t^r_j = 1 \mid t^s_i = 1\}$ becomes $p(t_j|t_i)$, and so on.

For a binary question, one may interpret the model as follows. Each respondent privately and independently conducts one toss of a biased coin, with unknown probability ω_H of heads. The result of the toss represents his opinion. Using this datum, he forms a posterior distribution, $p(\omega_H|t^r)$, whose expectation is the predicted frequency of heads. For example, if the prior is uniform, then the posterior distribution following the toss will be triangular on [0,1], skewed toward heads or tails depending on the result of the toss, with an expected value of one-third or two-thirds. However, if the prior is not uniform but strongly biased toward the opposite result (i.e., tails), then the expected frequency of heads following a heads toss might still be quite low. This would correspond to a prima facie unusual characteristic, such as having more than 20 sexual partners within the previous year.

An important simplification in the method is that I never elicit prior or posterior distributions, only answers and predicted frequencies. Denoting answers and predictions by $x^r = (x^r_1,..,x^r_m)$ ($x^r_k \in \{0,1\}$, $\Sigma_k x^r_k = 1$) and $y^r = (y^r_1,..,y^r_m)$ ($y^r_k \geq 0$, $\Sigma_k y^r_k = 1$), respectively, I calculate the population endorsement frequencies, $\bar{x}_k$, and the (geometric) average, $\bar{y}_k$, of predicted frequencies,

$$\bar{x}_k = \lim_{n\to\infty} \frac{1}{n}\sum_{r=1}^{n} x^r_k,$$

$$\log \bar{y}_k = \lim_{n\to\infty} \frac{1}{n}\sum_{r=1}^{n} \log y^r_k$$

Instead of applying a preset answer key, we evaluate answers according to their information score, which is the log-ratio of actual-to-predicted endorsement frequencies. The information score for answer k is

$$\log \frac{\bar{x}_k}{\bar{y}_k} \quad (1)$$

At least one answer will have a nonnegative information score. Variance in predictions tends to lower all $\bar{y}_k$ values and hence raises information scores.

The total score for a respondent combines the information score with a separate score for the accuracy of predictions (*20*):

$$\text{score for respondent } r = \text{information score} + \text{prediction score} = \sum_k x^r_k \log \frac{\bar{x}_k}{\bar{y}_k} + \alpha \sum_k \bar{x}_k \log \frac{y^r_k}{\bar{x}_k}, 0 < \alpha \quad (2)$$

Equation 2 is the complete payoff equation for the game. It is symmetric, and zero-sum if $\alpha = 1$. The first part of the equation selects a single information-score value, given that $x^r_k = 0$ for all answers except the one endorsed by r. The second part is a penalty proportional to the relative entropy (or Kullback-Leibler divergence) between the empirical distribution and r's prediction of that distribution (*21*, *22*). The best prediction score is zero, attained when prediction exactly matches reality, $y^r_k = \bar{x}_k$. Expected prediction score is maximized by reporting expected frequencies, $y^r_k = E\{\bar{x}_k|t^r\}$ (*2*). The constant α fine-tunes the weight given to prediction error.

To see how this works in the simple coin toss setting, imagine that there are only two equally likely possibilities: Either the coin is fair, or it is unfair, in which case it always comes up heads. A respondent who privately observes a single toss of tails knows that the coin is fair, and predicts a 50-50 split of observations. A respondent observing heads lowers the probability of fairness from the prior 1/2 to a posterior of 1/3, in accord with Bayes' rule, which in turn yields a predicted (i.e., expected) frequency of 1/6 for tails (multiplying 1/3 by 1/2). From the perspective of someone observing tails, the expectation of others' predictions of the frequency of tails will be a mix of predictions of 1/2 (from those tossing tails) and 1/6 (from those tossing heads), yielding a geometric mean clearly lower than his or her predicted frequency of 1/2. Hence, he or she expects that tails will prove to be more common than predicted and receive a positive information score. By contrast, heads is expected to be a surprisingly uncommon toss, because the predicted frequency of 1/2 is lower than the expectation of others' predictions, which is a mix of 1/2 and 5/6 predictions. A similar argument would show that those who draw heads should expect that heads will prove to be the answer with the high information score.

The example illustrates a general property of information scores. Namely, a truthful answer constitutes the best guess about the most surprisingly common answer, if "best" is defined precisely by expected information score and if other respondents are answering truthfully and giving truthful predicted frequencies. This property does not depend on the number of possible answers or on the prior (*23*). It leads directly to the equilibrium result [proof in the supporting online material (SOM) text].

For this theorem, assume that (i) every respondent r with opinion t^r forms a posterior over the population distribution of opinions, $p(\omega|t^r)$, by applying Bayes' rule to a common prior $p(\omega)$; (ii) $p(\omega|t^r) = p(\omega|t^s)$ if and only if $t^r = t^s$; and (iii) scores are computed according to Eq. 2. Then, (T1) truth-telling is a Nash equilibrium for any $\alpha > 0$: Truth-telling maximizes expected total score of every respondent who believes that others are responding truthfully; (T2) expected equilibrium information scores are nonnegative and attain a maximum for all respondents in the truth-telling equilibrium; (T3) for $\alpha = 1$, the game is zero-sum, and the total scores in the truth-telling equilibrium equal $\log p(\omega|t^r) + K$, with K set by the zero-sum constraint.

Truth-telling is defined as truthful answers, $x^r = t^r$, and truthful predictions, $y^r = E\{\omega|t^r\}$. T2 states that although there are other equilibria, constructed by mapping multiple true opinions into a single response category or by randomization, these less revealing equilibria result in lower information scores for all respondents. If needed, one can enhance the strategic advantage of truth-telling by giving relatively more weight to information score in Eq. 2 (*24*). For sufficiently small α, the expected total scores in the truth-telling equilibrium will Pareto-dominate expected scores in any other equilibrium. T3 shows that by setting $\alpha = 1$ we also have the option of presenting the survey as a purely competitive, zero-sum contest. Total scores then rank respondents according to how well they anticipate the true distribution of answers. Note that the scoring system asks only for

the expected distribution of true answers, $E\{\omega|t^r\}$ and not for the posterior distribution $p(\omega|t^r)$, which is an m-dimensional probability density function. Remarkably, one can infer which respondents assign more probability to the actual value of ω by means of a procedure that does not elicit these probabilities directly.

In previous economic research on incentive mechanisms, it has been standard to assume that the scorer (or the "center") knows the prior and posteriors and incorporates this knowledge into the scoring function (*7*–*9*, *25*). In principle, any change in the prior, whether caused by a change in question wording, in the composition of the sample, or by new public information, would require a recalculation of the scoring functions. By contrast, my method employs a universal "one-size-fits-all" scoring equation, which makes no mention of prior or posterior probabilities. This has three benefits for practical application. First, questions do not need to be limited to some pretested set for which empirically estimated base rates and conditional probabilities are available; instead, one can use the full resources of natural language to tailor a new set of questions for each application. Second, it is possible to apply the same survey to different populations, or in a dynamic setting (which is relevant to political polling). Third, one can honestly instruct respondents to refrain from speculating about the answers of others while formulating their own answer. Truthful answers are optimal for any prior, and there are no posted probabilities for them to consider, and perhaps reject.

These are decisive advantages when it comes to scoring complex, unique questions. In particular, one can apply the method to elicit honest probabilistic judgments about the truth value of any clearly stated proposition, even if actual truth is beyond reach and no prior is available. For example, a recent book, *Our Final Century*, by a noted British astronomer, gives the chances of human survival beyond the year 2100 at no better than 50:50 (*26*). It is a provocative assessment, which will not be put to the test anytime soon. With the present method, one could take the question: "Is this our final century?" and submit it to a sample of experts, who would each provide a subjective probability and also estimate probability distributions over others' probabilities. T1 implies that honest reporting of subjective probabilities would maximize expected information score. Experts would face comparable truth-telling incentives as if they were betting on the actual outcome [e.g., as in a futures market (*27*)] and that outcome could be determined in time for scoring.

I illustrate this with a discrete computation, which assumes that probabilities are elicited at 1% precision by means of a 100-point multiple-choice question (in practice, one would have fewer categories and smooth out the empirical frequencies). The population vector $\omega = (\omega_{00},..,\omega_{99})$ indexes the unknown distribution of such probabilities among experts. Given any prior, $p(\omega)$, it is a laborious but straightforward exercise to calculate expected information score as function of true personal probability and endorsed probability. Figure 1, lines A90 and B90, present the result of such calculations, with two different priors, $p_A(\omega)$ and $p_B(\omega)$, for experts who happen to agree that the probability of disaster striking before 2100 is 90%. The experts thus share the same assessment but have different theories about how their assessment is related to the assessment of others. Although lines A90 and B90 differ, the expected information score is in both cases maximized by a truthful endorsement of 90%. This confirms T1. In both cases, each expert believes that his subjective probability is pessimistic relative to the population: The expectation of others' probabilities, conditioned on a personal estimate of 90%, is only 65% with $p_A(\omega)$ and 54% with $p_B(\omega)$.

If the subjective probability shifts to 50%, the lines move to A50, B50, and the optimum, in both cases, relocates to 50%. Hence, the optimum automatically tracks changes in subjective belief, in this case the subjective probability of an unknown future event, but is invariant with respect to assumptions about how that belief is related to beliefs of other individuals. Changing these assumptions will simply lead back to the same recommendation: Truthfully report subjective probability.

Respondents are thus free to concentrate on their personal answer and need not worry about formulating an adequate prior. Any model of the prior is likely to be complex and involve strong assumptions. For example, in the calculations in Fig. 1, I assumed that experts' estimates are based on a private signal, distributed between zero and one, representing a personal assessment of the credibility of evidence supporting the bad outcome. The "credibility signal" is a valid but stochastic indicator of the true state of affairs: On the bad scenario, credibility signals are independent draws from a uniform distribution, so that some experts "get the message" and some do not; on the good scenario, they are independent draws from a triangular distribution, peaking at zero (no credibility) and declining linearly to one (full credibility). A prior probability of catastrophe then induces a monotonic mapping from credibility signals to posterior probabilities of catastrophe, as well as a prior over experts' probability estimates, $p(\omega)$.

Lines A and B differ in that the prior probability of catastrophe is presumed to be 50% for line A and 20% for line B. Expected scores are higher for B, because the 90% estimate is more surprising in that case.

One could question any of the assumptions of this model (*28*). However, changing the assumptions would not move the optimum, as long as the impersonally informative requirement is preserved. (The impersonally informative requirement means that two experts will estimate the same probability of catastrophe if and only if they share the same posterior distribution over other experts' probabilities). Thus, even though information scoring conditions success on the answers of other people, the respondent does not need to develop a theory of other people's answers; the most popular answer has no advantage of "winning," and the entire structure of mutual beliefs, as embodied in the prior, is irrelevant.

It is instructive to compare information scores with scores that would be computed if the scorer had a crystal ball and could score

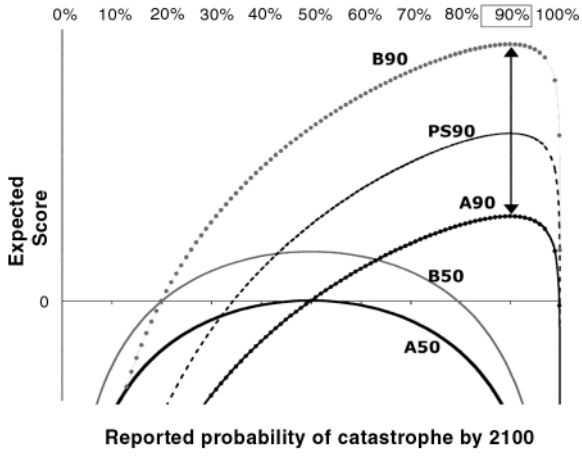


Fig. 1. The expected information score is maximized by a truthful report of subjective belief in a proposition (i.e., "this is our final century"), irrespective of priors (A or B) or subjective probability values (50% or 90%). Line A90 gives expected score for different reported probabilities when true personal estimate of catastrophe is 90% and prior probability is 50%. It is optimal to report 90% even though that is expected to be an unusually pessimistic estimate. Changing the prior to 20% (line B90) increases expected scores but does not displace the optimum. Changing subjective probability to 50% shifts the optimum to 50% (A50 assumes a 50% prior, B50 a 20% prior). Standard proper scoring (expectation of Eq. 3, displayed as line PS90) also maximally rewards a truthful report (90%). However, proper scoring requires knowledge of the true outcome, which may remain moot until 2100.

estimates for accuracy. The standard instrument for eliciting honest probabilities about publicly verifiable events is the logarithmic proper scoring rule (*2*, *4*, *29*). With the rule, an expert who announces a probability distribution $z = (z_1,..,z_n)$ over n mutually exclusive events would receive a score of

$$K + \log z_i \quad (3)$$

if event i is realized. For instance, an expert whose true subjective probability estimate that humanity will perish by 2100 is 90%, but who announced a possibly different probability z, would calculate an expected score of $0.9 \log z + 0.1 \log(1 - z)$, assuming, again, that there was some way to establish the true outcome. This expectation is maximized at the true value, $z = 0.90$, as shown by line PS90 in Fig. 1 (elevation is arbitrary). It is hard to distinguish proper scoring, which requires knowledge of the true outcome, from information scoring, which does not require such knowledge (*30*).

There are two generic ways in which the assumption of an impersonally informative prior might fail. First, a true answer might not be informative about population frequencies in the presence of public information about these frequencies (inducing a sharp prior). For instance, a person's gender would have minimal impact on their judgment of the proportion of men and women in the population. This would be a case of $t^r \neq t^s$ but $p(\omega|t^r) \cong p(\omega|t^s)$, and the difference between expected information scores for honest and deceptive answers would be virtually zero (though still positive). As shown below, the remedy is to combine the gender question with an opinion question that interacts with gender.

Second, respondents with different tastes or characteristics might choose the same answer for different reasons and hence form different posteriors. For example, someone with nonstandard political views might treat his or her liking for a candidate as evidence that most people will prefer someone else. This would be a case of: $p(\omega|t^r) \neq p(\omega|t^s)$ although $t^r = t^s$. Here, too, the remedy is to expand the questionnaire, allowing the person to reveal both the opinion and characteristic.

A last example, an art evaluation, illustrates both remedies. The example assumes existence of experts and laymen, and a binary state-of-nature: a question of whether a particular artist either does or does not represent an original talent. By hypothesis, art experts recognize this distinction quite well, but laymen discriminate poorly and, indeed, have a higher chance of enjoying a derivative artist than an original one. The fraction of experts is common knowledge, as are the other probabilities (Table 1).

In the short version of the survey, respondents only state their opinion; in the long version, they also report their expertise. Table 1 displays expected information scores for all possible answers, as a function of opinion and expertise. With the short version, truth-telling is optimal for experts but not for laymen, who do have a slight incentive to deceive if they happen to like the exhibition. With the long version, however, the diagonal, truth-telling entries have highest expected score. In particular, respondents will do better if they reveal their true expertise even though the distribution of expertise in the surveyed population is common knowledge.

Expected information scores in this and other examples reflect the amount of information associated with a particular opinion or characteristic. In Table 1, experts have a clear advantage even though they comprise a minority of the sample, because their opinion is more informative about population frequencies. In general, the expected information score for opinion i equals the expected relative entropy between distribution $p(\omega|t_k,t_i)$ and $p(\omega|t_k)$, averaged over all t_k. In words, the expected score for i is the information-theoretic measure of how much endorsing opinion i shifts others' posterior beliefs about the population distribution. An expert endorsement will cause greater shift in beliefs, because it is more informative about the underlying variables that drive opinions for both segments (*31*). This measure of impact is quite insensitive to the size of the expert segment or to the direction of association between expert and nonexpert opinion.

By establishing truth-telling incentives, I do not suggest that people are deceitful or unwilling to provide information without explicit financial payoffs. The concern, rather, is that the absence of external criteria can promote self-deception and false confidence even among the well-intentioned. A futurist, or an art critic, can comfortably spend a lifetime making judgments without the reality checks that confront a doctor, scientist, or business investor. In the absence of reality checks, it is tempting to grant special status to the prevailing consensus. The benefit of explicit scoring is precisely to counteract informal pressures to agree (or perhaps to "stand out" and disagree). Indeed, the mere existence of a truth-inducing scoring system provides methodological reassurance for social science, showing that subjective data can, if needed, be elicited by means of a process that is neither faith-based ("all answers are equally good") nor biased against the exceptional view.

Table 1. An incomplete question can create incentives for misrepresentation. The first pair of columns gives the conditional probabilities of liking the exhibition as function of originality (so that, for example, experts have a 70% chance of liking an original artist). It is common knowledge that 25% of the sample are experts, and that the prior probability of an original exhibition is 25%. The remaining columns display expected information scores. Answers with highest expected information score are shown by bold numbers. Truth-telling is optimal in the long version but not in the short version of the survey.

Opinion	Probability of opinion conditional on quality of exhibition		Expected score					
			Long version				Short version	
			Expert claim		Layman claim			
	Original	Derivative	Like	Dislike	Like	Dislike	Like	Dislike
Expert								
Like	70%	10%	+**575**	–776	–462	+67	+**191**	–57
Dislike	30%	90%	–934	+**95**	+84	–24	–86	+**18**
Layman								
Like	10%	20%	–826	+32	+**45**	–18	–66	+**12**
Dislike	90%	80%	–499	–156	–73	+**2**	–6	**–4**

References and Notes

1. C. F. Turner, E. Martin, Eds., *Surveying Subjective Phenomena* (Russell Sage Foundation, New York, 1984), vols. I and II.
2. R. M. Cooke, *Experts in Uncertainty* (Oxford Univ. Press, New York, 1991).
3. A formalized scoring rule has diverse uses: training, as in psychophysical experiments (*32*); communicating desired performance (*33*); enhancing motivation and effort (*34*); encouraging advance preparation, as in educational testing; attracting a larger and more representative pool of respondents; diagnosing suboptimal judgments (*4*); and identifying superior respondents.
4. R. Winkler, *J. Am. Stat. Assoc.* **64**, 1073 (1969).
5. W. H. Batchelder, A. K. Romney, *Psychometrika* **53**, 71 (1988).
6. In particular, this precludes the application of a futures markets (*27*) or a proper scoring rule (*29*).
7. C. d'Aspremont, L.-A. Gerard-Varet, *J. Public Econ.* **11**, 25 (1979).
8. S. J. Johnson, J. Pratt, R. J. Zeckhauser, *Econometrica* **58**, 873 (1990).
9. P. McAfee, P. Reny, *Econometrica* **60**, 395 (1992).
10. H. A. Linstone, M. Turoff, *The Delphi Method: Techniques and Applications* (Addison-Wesley, Reading, MA, 1975).
11. R. M. Dawes, in *Insights in Decision Making*, R. Hogarth, Ed. (Univ. of Chicago Press, Chicago, IL, 1990), pp. 179–199.
12. S. J. Hoch, *J. Pers. Soc. Psychol.* **53**, 221 (1987).
13. It is one of the most robust findings in experimental psychology that participants' self-reported characteristics—behavioral intentions, preferences, and beliefs—are positively correlated with their estimates of the relative frequency of these characteristics (*35*). The psychological literature initially regarded this as an egocentric error of judgment (a "false consensus") (*36*) and did not consider the Bayesian

explanation, as was pointed out by Dawes (*11*, *14*). There is still some dispute over whether the relationship is entirely consistent with Bayesian updating (*37*).
14. R. M. Dawes, *J. Exp. Soc. Psychol.* **25**, 1 (1989).
15. D. Fudenberg, J. Tirole, *Game Theory* (MIT Press, Cambridge, MA, 2000).
16. With finite players, the truth-telling result holds provided that the number of players exceeds some finite n, which in turn depends on $p(\omega)$.
17. J. M. Bernardo, A. F. M. Smith, *Bayesian Theory, Wiley Series in Probability and Statistics* (Wiley, New York, 2000).
18. R. J. Aumann, *Econometrica* **55**, 1 (1987).
19. More precisely, I assume in the SOM text that for any finite subset of respondents, there is a common and exchangeable prior over their opinions (hence, the prior is invariant under permutation of respondents). By de Finetti's representation theorem (*25*), this implies the existence of a probability distribution, $p(\omega)$, such that opinions are independent conditional on ω. Conditional independence ensures $t^r = t^s \Rightarrow p(\omega|t^r) = p(\omega|t^s)$. The reverse implication (i.e., that different opinions imply different posteriors) is also called stochastic relevance (*8*).
20. The finite n-player scoring formula ($n \geq 3$), for respondent r, is

$$\sum_{s\neq r}\sum_{k} x_k^r \log\frac{\bar{x}_k^{-rs}}{\bar{y}_k^{-rs}} + \alpha\sum_{s\neq r}\sum_{k}\bar{x}_k^{-rs}\log\frac{y_k^r}{\bar{x}_k^{-rs}}$$

where $\bar{x}_k^{-rs} = \left(\sum_{q\neq r,s} x_k^q + 1\right)/(n+m-2)$ and $\log\bar{y}_k^{-rs} = \sum_{q\neq r,s}\log y_k^q/(n-2)$. The score for r is built up from pairwise comparisons of r against all other respondents s, excluding from the pairwise calculations the answers and predictions of respondents r and s. To prevent infinite scores associated with zero frequencies, I replace the empirical frequencies with Laplace estimates derived from these frequencies. This is equivalent to "seeding" the empirical sample with one extra answer for each possible choice. Any distortion in incentives can be made arbitrarily small by increasing the number of respondents, n. The scoring is zero-sum when $\alpha = 1$.
21. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
22. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1954).
23. The key step in the proof involves calculation of expected information score for someone with personal opinion i but endorsing a possibly different answer j,

$$E\{\log\frac{\bar{x}_j}{\bar{y}_j}|t_i\} = \int_\Omega p(\omega|t_i)E\{\log\frac{\bar{x}_j}{\bar{y}_j}|\omega\}\,d\omega \quad (a)$$

$$= \int_\Omega p(\omega|t_i)\sum_{k=1}^{m}\omega_k\log\frac{\omega_j}{p(t_j|t_k)}d\omega \quad (b)$$

$$= \sum_{k=1}^{m} p(t_k|t_i)\int_\Omega p(\omega|t_k,t_i) \times \log\frac{p(t_j|\omega)}{p(t_j|t_k)}\frac{p(t_k|t_j,\omega)}{p(t_k|\omega)}d\omega \quad (c)$$

$$= \sum_{k=1}^{m} p(t_k|t_i)\int_\Omega p(\omega|t_k,t_i) \times \log\frac{p(\omega|t_k,t_j)}{p(\omega|t_k)}d\omega. \quad (d)$$

Once we reach (d), we can use the fact that the integral,

$$\int_\Omega p(\omega|t_k,t_i)\,\log p(\omega|t_k,t_j)\,d\omega,$$

is maximized when: $p(\omega|t_k,t_j) = p(\omega|t_k,t_i)$, to conclude that a truthful answer, i, will have higher expected information score than any other answer j. To derive (d), we first compute expected information score (a) with respect to the posterior distribution, $p(\omega|t_i)$, and use the assumption that others are responding truthfully to derive (b). For an infinite sample, truthful answers imply: $\bar{x}_j = \omega_j$, and truthful predictions: $\log\bar{y}_j = \Sigma_k\omega_k\log p(t_j|t_k)$, because the fraction ω_k of respondents who draw k will predict $p(t_j|t_k)$ for answer j. To derive (c) from (b), we apply conditional independence to write $\omega_k\ p(\omega|t_i)$ as $p(t_k|t_i)p(\omega|t_k,t_i)$, ω_j as $p(t_j|\omega)$, and 1 as $p(t_k|t_j,\omega)/p(t_k|\omega)$, which is inserted into the fraction. (d) follows from (c) by Bayes' rule.
24. The scoring system is not easy to circumvent by collective collusion, because if everyone agrees to give the same response then that response will no longer be surprisingly common, and will receive a zero information score. The prediction scores will also be zero in that case.
25. J. Cremer, R. P. McLean, *Econometrica* **56**, 1247 (1988).
26. M. Rees, *Our Final Century* (Heinemann, London, 2003).
27. J. Berg, R. Forsythe, T. A. Rietz, in *Understanding Strategic Interaction: Essays in the Honor of Reinhard Selten*, W. Albers, W. Guth, B. Hammerstein, B. Moldovanu, E. van Damme, Eds. (Springer, New York, 1997), pp. 441–463.
28. Certainly, the assumption of independent credibility signals is unrealistic in that it implies that expert opinion can in aggregate predict the true outcome perfectly; a more realistic model would have to interpose some uncertainty between the outcome and the totality of expert opinion.
29. L. J. Savage, *J. Am. Stat. Assoc.* **66**, 783 (1971).
30. Information scoring is nonmetric, and the 100 probability levels are treated as simply 100 distinct response categories. The smooth lines in Fig. 1 reflect smooth underlying priors, $p_A(\omega)$ and $p_B(\omega)$. Unlike proper scoring, information scoring could be applied to verbal expressions of probability ("likely," "impossible," etc.).
31. Precisely, if the opinions of one type of respondent are a statistical "garbling" of the opinions of a second type, then the first type will receive a lower score in the truth-telling equilibrium. Garbling means that the more informed individual could replicate the statistical properties of the signal received by the less informed individual, simply by applying a randomization device to his own signal (*38*).
32. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophyics* (Peninsula Publishing, Los Altos, CA, 1989).
33. W. Edwards, *Psychol. Rev.* **68**, 275 (1961).
34. C. F. Camerer, R. Hogarth, *J. Risk Uncert.* **18**, 7 (1999).
35. G. Marks, N. Miller, *Psychol. Bull.* **102**, 72 (1987).
36. L. Ross, D. Greene, P. House, *J. Exp. Soc. Psychol.* **13**, 279 (1977).
37. J. Krueger, R. W. Clement, *J. Pers. Soc. Psychol.* **67**, 596 (1994).
38. D. Blackwell, *Ann. Math. Stat.* **24**, 265 (1953).
39. I thank D. Mijovic-Prelec, S. Frederick, D. Fudenberg, J. R. Hauser, M. Kearns, E. Kugelberg, R. D. Luce, D. McAdams, S. Seung, R. Weaver, and B. Wernerfelt for comments and criticism. I acknowledge early support for this research direction by Harvard Society of Fellows, MIT E-Business Center, and the MIT Center for Innovation in Product Development.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5695/462/DC1
SOM Text

28 June 2004; accepted 15 September 2004

Single-Atom Spin-Flip Spectroscopy

A. J. Heinrich,* J. A. Gupta, C. P. Lutz, D. M. Eigler

We demonstrate the ability to measure the energy required to flip the spin of single adsorbed atoms. A low-temperature, high–magnetic field scanning tunneling microscope was used to measure the spin excitation spectra of individual manganese atoms adsorbed on Al_2O_3 islands on a NiAl surface. We find pronounced variations of the spin-flip spectra for manganese atoms in different local environments.

The magnetic properties of nanometer-scale structures are of fundamental interest and may play a role in future technologies, including classical and quantum computation. Such magnetic structures are composed of magnetic atoms in precise arrangements. The magnetic properties of each atom are profoundly influenced by its local environment. Magnetic properties of atoms in a solid can be probed by placing the atoms in tunnel junctions. Early experiments with planar metal-oxide-metal tunnel junctions doped with paramagnetic impurities exhibited surprisingly complex conductance spectra described as "zero-bias anomalies" (*1*–*4*). Such anomalies were shown to reflect both spin-flips driven by inelastic electron scattering and Kondo interactions of magnetic impurities with tunneling electrons (*5*–*7*). Single, albeit unknown, magnetic impurities were later studied in nanoscopic tunnel junctions (*8*, *9*). Recently, magnetic properties of single-molecule transistors that incorporated either one or two magnetic atoms were probed by means of their elastic conductance spectra (*10*, *11*). These measurements determined g values and showed field-split Kondo resonances due to the embedded magnetic atoms.

The scanning tunneling microscope (STM) offers the ability to study single magnetic moments in a precisely characterized local environment and to probe the variations in magnetic properties with atomic-scale spatial resolution. Previous STM studies of atomic-scale magnetism include Kondo resonances of magnetic atoms on metal surfaces (*12*, *13*), increased noise at the Larmor frequency (*14*, *15*),

IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA.

*To whom correspondence should be addressed. E-mail: heinrich@almaden.ibm.com

and spin-polarized tunneling (*16*). We demonstrate a technique for measuring the spin excitation spectra of individual atoms adsorbed on a surface using inelastic electron tunneling spectroscopy (IETS) with a STM. Combined with the STM's capability to fabricate, image, and modify atomically precise structures, this technique provides a powerful new tool for studying and engineering the local magnetic properties of nanometer-scale systems.

IETS measures excitation energies, such as vibrational energies, of atoms or molecules within tunnel junctions (*17*, *18*). Above a threshold voltage, electrons are able to transfer energy to these excitations during the tunneling process. This additional tunneling channel results in an upward step in conductance at the threshold voltage. For the measurements reported here, tunneling electrons lose energy to spin-flip excitations of single Mn atoms. The signature of Zeeman splitting in spin-flip IETS is a step up in conductance at an energy proportional to the applied magnetic field.

We used a home-built, ultrahigh-vacuum STM that reaches a base temperature of 0.6 K by means of a single-shot pumped ^{3}He refrigerator. The STM is vibrationally isolated and at the same time thermally coupled to the ^{3}He liquid by suspending the STM chamber directly above the liquid. We liquefied the ^{3}He using the Joule-Thomson effect, obviating the need for a pumped ^{4}He reservoir. Magnetic fields up to $B = 7$ T were applied in the plane of the sample. NiAl(110) samples were prepared in vacuum by repeated sputter/anneal cycles. Samples were then exposed to ~10 Langmuir of O_2 at ~500 K and further annealed at 1200 K. This resulted in the growth of patches of Al_2O_3 two layers thick (0.5 nm) interspersed with regions of bare NiAl (*19*). Samples were then transferred into the STM, and Mn atoms were subsequently evaporated onto the cold surface. Mn has partially filled *d*-orbitals, and the free atom has a total spin of $S = 5/2$. The differential conductance, dI/dV, was measured using lock-in detection of the tunnel current I by adding a 50-μV root mean square modulation at 829 Hz to the sample bias voltage V.

A topograph of the partially oxidized NiAl surface (Fig. 1A) shows that the bare metal and the Al_2O_3 oxide regions are atomically flat. Contrast on the metal is caused by standing waves in surface-state electrons (*20*). The oxide has a nearly rectangular unit cell 1.06 nm by 1.79 nm, which yields a complex but nearly periodic pattern in the STM topograph (*21*). The cold sample was subsequently dosed with a small amount of Mn, and the same area was imaged again (Fig. 1B). Single Mn atoms are seen as protrusions with an apparent height of 0.13 nm on the bare metal surface and 0.16 nm on the oxide. The density of Mn atoms on the oxide is significantly smaller than on the metal, presumably due to a lower sticking probability and motion along the oxide surface during adsorption (*22*).

The upper set of spectra in Fig. 1C shows the marked magnetic-field dependence of the conductance when the tip is positioned over a Mn atom on the oxide. At $B = 7$ T, the conductance is reduced near zero bias, with symmetric steps up to a ~20% higher conductance at an energy of $|\Delta| \sim 0.8$ meV. These conductance steps are absent at $B = 0$. Furthermore, no conductance steps are observed when the tip is positioned over the bare oxide surface, over the bare metal surface, or over a Mn atom on the metal surface (Fig. 1, C and D). We verified that these conductance spectra are characteristic for single Mn atoms on oxide terraces and on bare NiAl(110) by measuring many Mn atoms with different atomic arrangements at the STM tip apex. The spatial extent of the conductance step can be visualized by measuring dI/dV while imaging the Mn atom (Fig. 1, E and F). We find that the dI/dV signal is localized to an area 1 nm in diameter, comparable to the atom's apparent lateral extent in the corresponding STM topograph.

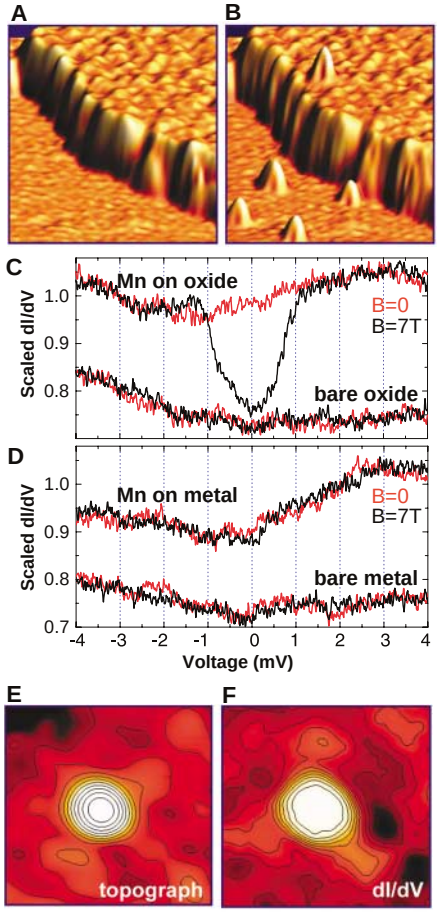


Fig. 1. Comparison of Mn atoms on oxide and on metal. (**A**) STM constant-current topograph of a NiAl (110) surface partially covered with Al_2O_3 (upper right). Image: 20 nm by 10 nm; V = 100 mV, I = 50 pA. (**B**) Same area after dosing with Mn. (**C**) Conductance spectra at $T = 0.6$ K on the Mn atom on oxide (upper curves) measured at $B = 7$ T (black) and $B = 0$ T (red). The lower curves (shifted for clarity) were measured over the bare oxide surface. (**D**) Conductance spectra on a Mn atom on NiAl (upper curves) and on the bare NiAl surface (lower curves). All spectra in (C) and (D) were acquired with a nominal conductance of 10 nA/V (I = 50 pA at V = 5 mV) and normalized to unity for $|V| > 2$ mV to emphasize differences in low-bias features. (**E**) Topograph of the Mn atom on oxide. Image: 2.8 nm by 2.8 nm; $B = 7$ T, $T = 0.6$ K, $V = 2$ mV, $I = 20$ pA, $V_{AC} = 0.5$ mV_{rms}. (**F**) Spatial map of dI/dV acquired concurrently; an increased signal (light area) maps the spatial extent of the spin-flip conductance step.

The characteristic signature of spin-flip IETS is a step up in the differential conductance dI/dV at a bias voltage corresponding to the Zeeman energy $\Delta = g\mu_B B$, where μ_B = 57.9 μeV/T is the Bohr magneton and g = 2.0023 for a free electron. Figure 2A shows that the conductance step shifts to higher energy with increased field. Broadening of this step is due mainly to the effect of temperature, with contributions from the ac voltage modulation, spin lifetime, and instrumental noise. The thermal broadening of tip and sample densities of states can be calculated by twice convolving an intrinsically sharp step with the

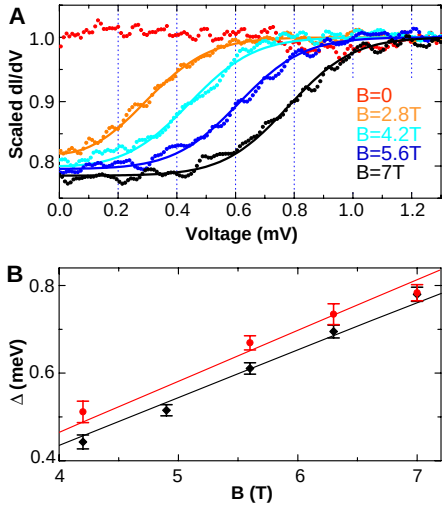


Fig. 2. Shift of the spin-flip conductance step with magnetic field. (**A**) Conductance spectra (points) for an isolated Mn atom on oxide at different magnetic fields. Solid lines show fits to the temperature-broadened step model (see text). The data fit well to a step height of 20.5% for all fits except the highest field, where a slight tip modification changed the step height to 21.5%. The effective temperature in all curves was $T = 0.85$ K. All spectra were acquired with a nominal conductance of 10 nA/V (I = 50 pA at V = 5 mV) and normalized to unity for large $|V|$ (see text). (**B**) Magnetic field dependence of the Zeeman energy Δ. Black points are extracted from the fits in (A), and red points were taken on a Mn atom near the edge of an oxide patch. Linear fits (black and red lines) constrained to $\Delta = 0$ at $B = 0$ yield g values of 1.88 and 2.01, respectively.

derivative of the Fermi-Dirac distribution (*23*). To fit our experimental data, we combine temperature with other sources of broadening by using an effective temperature (*24*). From the fits we extract the Zeeman splitting Δ for each value of magnetic field. We display the measured *dI*/*dV* curves in Fig. 2A by normalizing them to unity for voltages outside of the spin-flip region, by using the fit to establish the large-voltage conductance.

The measured Zeeman splitting is proportional to the magnetic field (Fig. 2B, black points). The data fit well with a straight line through the origin and a slope that corresponds to $g = 1.88 \pm 0.02$ (*25*). A different Mn atom, this one within 1 nm of the edge of an oxide patch, shows a significantly different g value (red points) of $g = 2.01 \pm 0.03$. The only difference between these two Mn atoms is the local environment: They have different lateral distances to bare metal region; they may sit at different binding sites in the oxide unit cell;

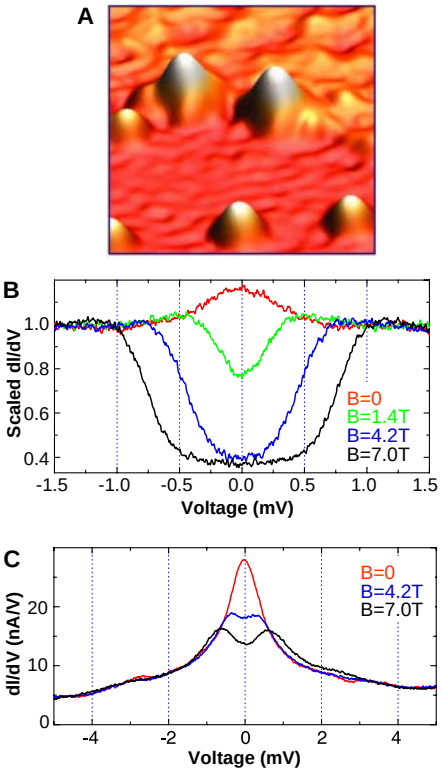


Fig. 3. Conductance spectra of Mn showing Kondo resonances. (**A**) Topograph (6 nm by 10 nm) of Mn atoms bound near the interface between an oxide patch (upper half) and bare metal. (**B**) Conductance spectra of the Mn atom at top center in (A). The zero-field spectrum shows a Kondo resonance: a Lorentzian-shaped rise in conductance near $V = 0$. At high fields a large spin-flip step dominates the Kondo signal. (**C**) Kondo resonance for a Mn atom (or cluster) near the boundary of an oxide patch that appears as a 0.17-nm-high protrusion (not pictured). The peak represents a factor-of-5 rise in conductance at zero bias. The Kondo peak broadens and splits symmetrically at higher magnetic fields.

and perhaps more importantly, we expect the oxide patch to show reconstruction near the boundary to minimize its energy. We are not aware of any studies of the detailed atomic structure of the oxide patches near their boundaries. We have verified that the values we measure for Δ are insensitive to the height and lateral position of the tip, indicating that the tip serves as a nonperturbative probe of the spin properties of the adsorbed atom.

Mn atoms on the oxide that are laterally near metal-oxide interfaces (e.g., Fig. 3A) can exhibit spectra that are markedly different from those of the isolated Mn atoms on oxide terraces. Both of the interfacial Mn atoms in Fig. 3A have an apparent height similar to that of single Mn on oxide terraces. Whereas the left Mn atom shows an IETS spectrum similar to those in Figs. 1 and 2, the right Mn atom shows much larger (~60%) steps in conductance (Fig. 3B). In addition, Fig. 3B shows a peak at zero field and zero bias. This peak splits and shifts to higher energy as the magnetic field is increased. Although the magnitude decreases sharply with field, the peaks remain clearly present at all magnetic fields. This behavior agrees well with the perturbation theory that was developed in the context of planar tunneling devices (*5*).

The relative strength of the zero-bias conductance peak and spin-flip steps can vary dramatically. Figure 3C shows an example of a Mn atom (or cluster) where the conductance peak dominates the spin-flip steps. At zero field, the zero-bias conductance is enhanced by a factor of ~5 relative to the background. The peak splits with magnetic field, and no clear spin-flip steps are observed.

The spectra in Fig. 3, B and C, show the hallmarks of a Kondo resonance: a narrow conductance peak with Lorentzian shape at zero bias that splits with magnetic field. The Kondo effect reflects the spin-flip interactions of conduction electrons with a localized magnetic impurity. The full width at half maximum of the resonance can be used to extract a Kondo temperature of $T_K \sim 3$ K in Fig. 3B, and $T_K \sim 6$ K in Fig. 3C (*26*, *27*). The enhanced zero-bias conductance seen here is a simpler manifestation of Kondo physics than obtained in previous STM studies, where a more complicated Fano line shape for Kondo resonances reflected interference effects (*12*, *13*, *26*, *27*). Unlike these earlier STM studies, where magnetic atoms were directly adsorbed on a metal surface, here the interaction between the Mn atom and the NiAl conduction electrons is mediated by an oxide film. We note that Mn adsorbed directly on NiAl does not show any Kondo signature in the 1- to 100-meV energy range studied here.

The zero-bias conductance peak for Mn on Al_2O_3 is comparable to Kondo effects observed in other nanostructures (*9*–*11*). However, the device characteristics in these nanostructures varied considerably, due presumably to uncontrolled variations in the molecular conformation, binding sites, electrode structures, and neighboring molecules. It is one of the strengths of STM to be able to characterize and control each of these variables.

The zero-bias anomaly in thin-film tunnel junctions showed a spin-flip channel with inhomogeneous broadening that was much larger than the sample temperature. This broad linewidth was attributed in part to the spatial average over impurities in the junction with differing g values (*3*, *4*). Our observations indicate that such spatially averaged studies may reflect not only different g values but also site-dependent amplitudes for the spin-flip and Kondo channels.

The ability to directly measure the g value of individual atoms with the STM enables site-specific study of magnetic moments. When combined with the STM's capability to assemble atomically precise structures, spin excitations can now be studied in custom-engineered nanostructures. If atoms with spins can be coupled to each other in a controlled fashion, it might be possible to use the spin degree of freedom to transmit and process information on the atomic scale (*28*).

References and Notes

1. A. F. G. Wyatt, *Phys. Rev. Lett.* **13**, 401 (1964).
2. L. Y. L. Shen, J. M. Rowell, *Phys. Rev.* **165**, 566 (1968).
3. R. H. Wallis, A. F. G. Wyatt, *J. Phys. C* **7**, 1293 (1974).
4. S. Bermon, D. E. Paraskevopoulos, P. M. Tedrow, *Phys. Rev. B* **17**, 2110 (1978).
5. J. A. Appelbaum, *Phys. Rev.* **154**, 633 (1967).
6. P. W. Anderson, *Phys. Rev. Lett.* **17**, 95 (1966).
7. A. Rosch, J. Paaske, J. Kroha, P. Wölfle, *Phys. Rev. Lett.* **90**, 076804 (2003).
8. S. Gregory, *Phys. Rev. Lett.* **68**, 2070 (1992).
9. D. C. Ralph, R. A. Buhrman, *Phys. Rev. Lett.* **72**, 3401 (1994).
10. J. Park *et al.*, *Nature* **417**, 722 (2002).
11. W. Liang, M. P. Shores, M. Bockrath, J. R. Long, H. Park, *Nature* **417**, 725 (2002).
12. J. Li, W.-D. Schneider, R. Berndt, B. Delley, *Phys. Rev. Lett.* **80**, 2893 (1998).
13. V. Madhavan, W. Chen, T. Jamneala, M. F. Crommie, N. S. Wingreen, *Science* **280**, 567 (1998).
14. Y. Manassen, R. J. Hamers, J. E. Demuth, A. J. Castellano Jr., *Phys. Rev. Lett.* **62**, 2531 (1989).
15. C. Durkan, M. E. Welland, *Appl. Phys. Lett.* **80**, 458 (2002).
16. S. Heinze *et al.*, *Science* **288**, 1805 (2000).
17. P. K. Hansma, Ed., *Tunneling Spectroscopy: Capabilities, Applications, and New Techniques* (Plenum, New York, 1982).
18. B. C. Stipe, M. A. Rezaei, W. Ho, *Science* **280**, 1732 (1998).
19. A. Stierle *et al.*, *Science* **303**, 1652 (2004).
20. Z. Song, J. I. Pascual, H. Conrad, K. Horn, H.-P. Rust, *Surf. Sci.* **491**, 39 (2001).
21. M. Kulawik, N. Nilius, H.-P. Rust, H.-J. Freund, *Phys. Rev. Lett.* **91**, 256101 (2003).
22. N. Nilius, T. M. Wallis, W. Ho, *Phys. Rev. Lett.* **90**, 046808 (2003).
23. A. Kogan *et al.*, http://xxx.lanl.gov/abs/cond-mat/0312186 (2003).
24. The best fit to the measured curves gives an effective temperature of 0.85 K. This is slightly higher than the actual temperature of the STM (0.6 K), which was measured by fitting *dI*/*dV* curves of a superconducting sample. This higher effective temperature is due mainly to residual high-frequency (RF) voltage noise.
25. The standard deviation uncertainties quoted here are derived from the uncertainty in the fits at each field (error bars in Fig. 2B). These uncertainties come from

instrumental noise, and from background spectral features due to the tip and surface that happen to lie near the Zeeman energy. For simplicity we have chosen not to subtract these background spectral features before fitting the data.
26. K. Nagaoka, T. Jamneala, M. Grobis, M. F. Crommie, *Phys. Rev. Lett.* **88**, 077205 (2002).
27. N. Knorr, M. A. Schneider, L. Diekhöner, P. Wahl, K. Kern, *Phys. Rev. Lett.* **88**, 096804 (2002).
28. A. J. Heinrich, C. P. Lutz, J. A. Gupta, D. M. Eigler, *Science* **298**, 1381 (2002).
29. We gratefully acknowledge help with the construction of the STM from B. Melior and helpful discussions with J. Kroha, B. A. Jones, A. H. Castro-Neto, A. V. Balatsky, and D. Loss. This work was supported in part by the Defense Advanced Research Projects Agency.

4 June 2004; accepted 23 August 2004
Published online 9 September 2004;
10.1126/science.1101077
Include this information when citing this paper.

Stable Low-Pressure Hydrogen Clusters Stored in a Binary Clathrate Hydrate

Louw J. Florusse,[1] Cor J. Peters,[1] Joop Schoonman,[2] Keith C. Hester,[3] Carolyn A. Koh,[3] Steven F. Dec,[3] Kenneth N. Marsh,[4] E. Dendy Sloan[3]*

Thermodynamic, x-ray diffraction, and Raman and nuclear magnetic resonance spectroscopy measurements show that clusters of H_2 can be stabilized and stored at low pressures in a sII binary clathrate hydrate. Clusters of H_2 molecules occupy small water cages, whereas large water cages are singly occupied by tetrahydrofuran. The presence of this second guest component stabilizes the clathrate at pressures of 5 megapascals at 279.6 kelvin, versus 300 megapascals at 280 kelvin for pure H_2 hydrate.

Clathrate hydrates are crystalline inclusion compounds composed of a hydrogen-bonded water host lattice and one or more types of guest molecules. The two most common clathrate hydrate structures are sI and sII. sI hydrate has two small 5^{12} cages and six large $5^{12}6^2$ cages per unit cell. sII hydrate has sixteen 5^{12} cages and eight $5^{12}6^4$ cages per unit cell (*1*). Hydrogen clusters can be stabilized in a clathrate hydrate at extremely high pressures (typically 220 MPa at 249 K) (*2*, *3*). It has been suggested that in the presence of a second guest component, hydrogen is excluded from the clathrate framework (*4*, *5*) or is included to a minor extent (*6*). Here we show that hydrogen clusters can be stabilized and stored at low pressures within the clathrate hydrate lattice by stabilizing the large water cages with a second guest component, tetrahydrofuran (THF). Storage of hydrogen in molecular form within a water clathrate framework at low pressures and near-ambient temperatures could achieve an increase in the overall energy efficiency (hydrogen is stored in molecular form without the need of a chemical reaction for its release), environmental cleanliness, and safety of hydrogen storage as compared to pure hydrogen hydrate or metal hydrides.

We investigated the phase behavior of the system $H_2O + H_2 + THF$ using a Cailletet facility for pressures up to 15 MPa and an autoclave facility for pressures up to 100 MPa (*7*–*9*). THF was selected as the second guest component because it is completely miscible with water and THF + water forms a sII hydrate alone. To gain insight into the effect of the promoter (THF) guest molecule on the phase behavior of the binary system $H_2O + H_2$, we measured the hydrate equilibrium of the binary system H_2O + THF. The results demonstrate that H_2/THF hydrate is stable at lower pressures than pure hydrogen hydrate and higher temperatures than pure THF hydrate (Fig. 1).

X-ray powder diffraction (XRPD) data show that the H_2/THF hydrate has a sII hydrate crystal structure (*1*), which is the same structure as reported by Mao *et al.* (*2*, *3*) for pure H_2 hydrate. The XRPD pattern in Fig. 2 can be indexed with a face-centered cubic unit cell (space group $Fd3m$) with $a = 17.225 \pm 0.019$ Å. This agrees with the reported value ($a = 17.236$ Å) for the lattice parameter of sII pure THF hydrate (*10*). The formation of the H_2/THF clathrate structure was also verified with Raman spectroscopy. The intramolecular vibrational modes of THF molecules are observed between 850 and 975 cm^{-1}. These modes consist primarily of C-C-C-C stretching motions (*11*). The spectral feature due to this group of modes is termed the ring breathing mode (*11*). In H_2/THF hydrate, the THF ring breathing mode appears as a single peak at around 920 cm^{-1}, which is different from the corresponding H_2/THF/D_2O liquid spectrum (Fig. 3, left). In the liquid, THF molecules hydrogen bond with D_2O molecules. Hence, in the liquid two peaks are observed at 890 cm^{-1} (resulting from the symmetric C-O-C stretching vibration) and 920 cm^{-1} (*12*).

We confirmed the inclusion of hydrogen molecules within the clathrate framework using Raman and solid-state magic-angle spinning (MAS) nuclear magnetic resonance (NMR) spectroscopy (*9*). The appearance of hydrogen roton peaks $S_0(0)$, $S_0(1)$, $S_0(2)$, and $S_0(J)$ represents pure rotational excitations (where the rotational quantum number $J = 0$, 1, and 2, respectively) at 300 to 850 cm^{-1} (Fig. 3, left). These rotons are similar in appearance to pure hydrogen and more intense than those seen in the H_2/THF/D_2O liquid. This result indicates that hydrogen molecules incorporated within the clathrate cages are still in free rotational states. The hydrogen molecules remain unbonded to each other or to water in the clathrate.

A single broad H-H vibron peak [pure molecular vibration excitation, $Q_{\Delta\nu}(J)$, where $\Delta\nu$ is the difference between the final and

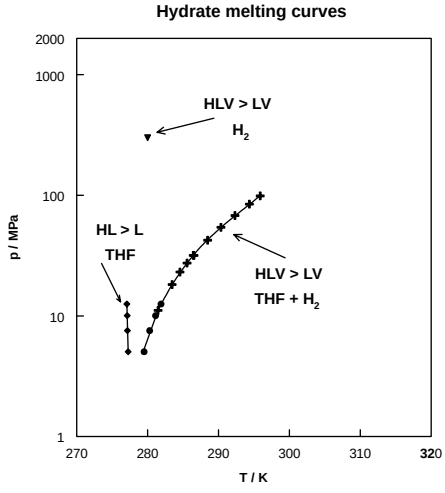


Fig. 1. Comparison of the phase transition HLV → LV in the ternary system $H_2O + H_2$ + THF with the phase transition HL → L in the binary system H_2O + THF. Diamonds and solid circles are data obtained with the Cailletet facility, and the crosses are data measured with the autoclave facility. The full curve represents the best fit to the experimental data. THF hydrate or THF/H_2 hydrate is only stable in the regions above and to the left of the two-phase or three-phase equilibrium line, respectively. The conditions for stabilizing pure H_2 hydrate (*3*) are represented by a solid triangle.

[1]Faculty of Applied Sciences, Department of Chemical Technology, Physical Chemistry and Molecular Thermodynamics; [2]Laboratory for Inorganic Chemistry, Delft Institute for Sustainable Energy; Delft University of Technology, Julianalaan 136, 2628 BL Delft, Netherlands. [3]Center for Hydrate Research, Department of Chemical Engineering, Colorado School of Mines, Golden, CO 80401, USA. [4]Department of Chemical and Process Engineering, University of Canterbury, Christchurch, Private Bag 4800, New Zealand.

*To whom correspondence should be addressed. E-mail: esloan@mines.edu

initial vibrational quantum levels] at around 4125 cm^{-1} is seen in the H_2/THF clathrate (Fig. 3, right). In contrast, the vibron peak is weak in the spectrum for H_2/THF/D_2O liquid. The vibron peak of the hydrate is different from that obtained for other known phases of H_2-H_2O (*13*). It should be noted that the H_2-H_2O phase referred to as a "clathrate hydrate" by Dyadin *et al*. (*14*) is based on an ice II framework rather than the sII clathrate hydrate. The H-H vibron peaks of hydrogen molecules incorporated in ice Ic (*13*), ice II (*13*), and ice Ih (*15*) at high pressures are at higher frequencies than those of pure hydrogen. In agreement with spectra reported by Mao *et al*. (*2*, *3*), the H-H vibron peak of the H_2/THF hydrate is below the dominant $Q_1(1)$ vibron peak at 4155 cm^{-1} observed in H_2-ice systems (*13*). The $Q_1(1)$ vibron peak in H_2-ice is about 10 to 20 cm^{-1} higher than that in pure H_2 (*12*). Mao *et al*. attributed the lower shift (or "softening") in the H-H vibron peak of pure H_2 clathrate to the increased intermolecular vibrational coupling within the H_2 clusters and a substantial gas-to-crystal frequency shift (*2*, *3*).

The single broad H-H vibron peak we observed in H_2/THF hydrate is different from the vibron spectrum recorded by Mao *et al*. for pure H_2 hydrate (*2*, *3*). The vibron peaks reported by Mao *et al*. (*2*, *3*) include two multiplet groups of equal intensity. They attributed the lower frequency group at 4115 to 4135 cm^{-1} to the loosely fitted molecular cluster of H_2 in a $5^{12}6^4$ large cage, whereas the higher frequency group at 4135 to 4155 cm^{-1} was assigned to the bimolecular cluster in a 5^{12} small cage. In the case of H_2/THF hydrate, we observed a broad H-H vibron peak at 4125 cm^{-1} (vibron peak shifts are pressure-dependent; Q_1 decreases with increasing pressure). In this clathrate structure, it is expected that all or most of the large cages are filled with THF, leaving the small cages available to store hydrogen. In view of this, we suggest that the peak at around 4125 cm^{-1} is attributed to H_2 molecules occupying a small cage. It may be that the vibron frequencies are affected by the size of the hydrogen cluster in particular clathrate cages, which would also affect the extent of intermolecular interactions between guest and host molecules.

Solid-state ^{1}H MAS NMR spectra of H_2/tetrahydrofuran-d_8 (H_2/TDF) hydrate (Fig. 4A) show one broad, intense resonance line with a chemical shift of 4.3 parts per million (ppm) and two lower intensity, sharper resonance lines with chemical shifts of 1.5 and 3.3 ppm. Structural assignments of these three resonance lines are readily made on the basis of comparison with the ^{1}H MAS NMR spectrum of TDF hydrate (Fig. 4B). The peaks with chemical shifts of 1.5 and 3.3 ppm in both ^{1}H MAS NMR spectra are due to residual–CHD group protons in the two nonequivalent sites of the TDF molecule. The peak with a chemical shift of 6.5 ppm in the TDF hydrate spectrum is due to residual HDO protons (due to D_2O with a small amount of HDO impurity) in the hydrate lattice. Thus, the broad, intense resonance line at 4.3 ppm in the H_2/TDF hydrate is due to H_2 molecules encapsulated in this sII hydrate. The H_2 molecules most likely occupy only the small cages to a significant extent because the TDF molecules occupy nearly all the large cavities in this sII hydrate.

Although the ^{1}H MAS NMR spectrum of the H_2/TDF hydrate shows no definitive evidence of multiple H_2 gas occupancy of a single small cage, the relatively large linewidth observed for its resonance line is consistent with an increase of homonuclear dipolar broadening that most likely would occur if each small cage contained more than one H_2 molecule guest. However, integrating

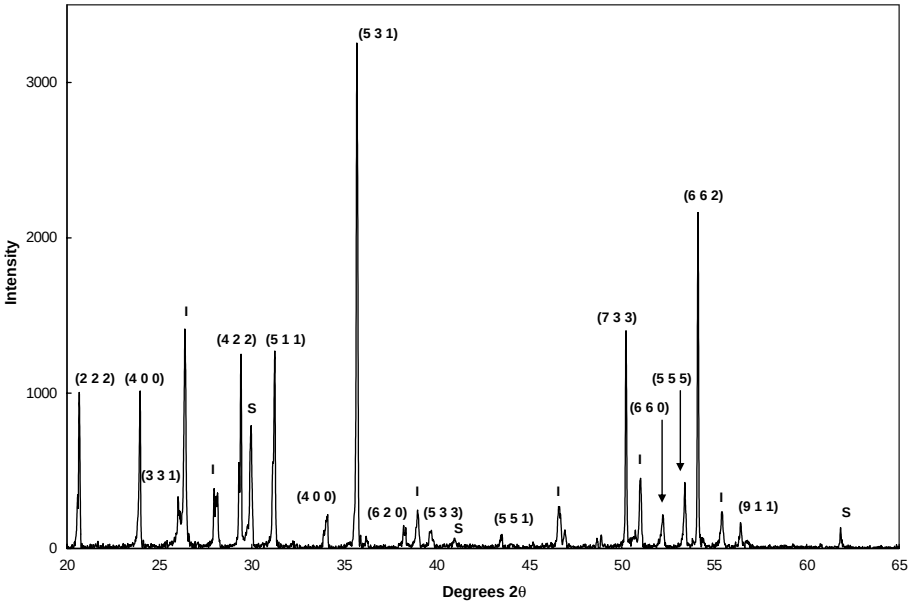


Fig. 2. XRPD pattern of H_2/THF hydrate at 190 K and atmospheric pressure.

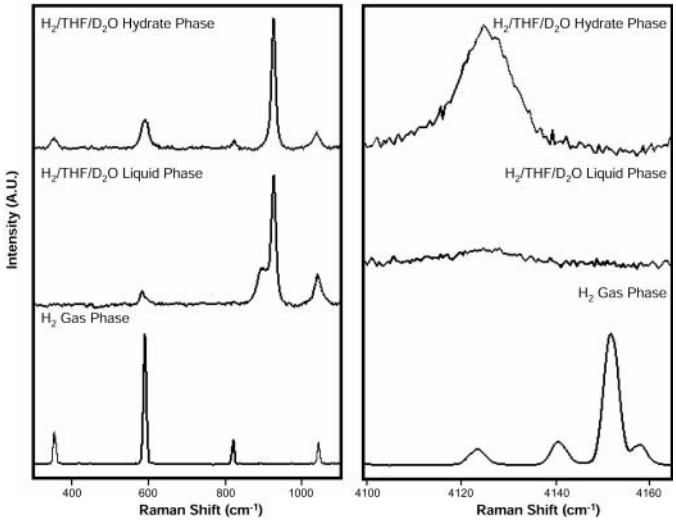


Fig. 3. Raman spectra of H_2/THF/D_2O hydrate compared to H_2/THF/D_2O liquid and pure H_2 gas of (**left**) the hydrogen molecular rotons $S_o(0)$, $S_o(1)$, and $S_o(2)$ and (**right**) the H-H vibrons.

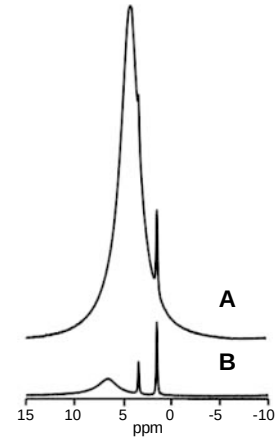


Fig. 4. ^{1}H MAS NMR spectra of (**A**) H_2/TDF/D_2O hydrate and (**B**) TDF/D_2O hydrate.

the H_2 peak in the MAS NMR spectrum gives an occupancy of about 0.5 H_2 molecules per small cage on average. In contrast, estimates of occupancy from volumetric measurements performed during hydrate dissociation give on average 1.0 molecules of H_2 per small cage. One explanation for this discrepancy between NMR and volumetric measurements is that there was a loss of H_2 during NMR measurements, because the MAS sample rotors were open to the atmosphere. This explanation is supported by repeat NMR experiments that show that the size of the H_2 peak can vary among samples. Based on the NMR line width of H_2, some small cages may be doubly occupied, whereas other small cages are singly occupied.

Our results show that promoter guest molecules can be used to store hydrogen in a binary clathrate hydrate at low pressures. Storage capacities might be increased by optimization of the promoter system. In the case of sII binary hydrogen hydrate, with double occupancy of the small cavities by H_2 and the large cavities partially occupied by THF, the mass of hydrogen could be up to 4%.

References and Notes

1. E. D. Sloan, *Clathrate Hydrates of Natural Gas* (Marcel Dekker, New York, ed. 2, 1998).
2. W. L. Mao *et al.*, *Science* **297**, 2247 (2002).
3. W. L. Mao, H. K. Mao, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 708 (2004).
4. G. D. Holder *et al.*, *Ind. Eng. Chem. Process Des. Dev.* **22**, 170 (1983).
5. S.-X. Zhang, G.-J. Chen, C.-F. Ma, L.-Y. Yang, T.-M. Guo, *J. Chem. Eng. Data* **45**, 908 (2000).
6. R. M. Barrer, D. J. Ruzicka, *Trans. Faraday Soc.* **58**, 2239 (1962).
7. S. Raeissi, C. J. Peters, *J. Supercritical Fluids* **20**, 221 (2001).
8. T. W. Loos, A. J. M. de Wijen, G. A. M. Diepen, *J. Chem. Thermodyn.* **12**, 193 (1980).
9. Materials and methods are available as supporting material on *Science* Online.
10. J. S. Tse, *J. Physique* **48 C1**, 543 (1987).
11. C. A. Tulk, D. D. Klug, J. A. Ripmeester, *J. Phys. Chem. A* **102**, 8743 (1998).
12. M. Zugik, thesis, King's College, London (2001).
13. W. L. Vos, L. W. Finger, R. J. Hemley, H. K. Mao, *Phys. Rev. Lett.* **71**, 3150 (1993).
14. Y. A. Dyadin *et al.*, *Mendeleev Commun.* **9**, 209 (1999).
15. S. A. Sandford, L. J. Allamandola, T. R. Geballe, *Science* **262**, 400 (1993).

16. We thank J. Ivanic, T. Strobel, and C. Taylor for their help with the volumetric measurements and K. Miller for his advice on state-of-the-art hydrogen storage.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5695/469/DC1
Materials and Methods

28 June 2004; accepted 10 September 2004

Mass Spectrometry Sampling Under Ambient Conditions with Desorption Electrospray Ionization

Zoltán Takáts, Justin M. Wiseman, Bogdan Gologan, R. Graham Cooks*

A new method of desorption ionization is described and applied to the ionization of various compounds, including peptides and proteins present on metal, polymer, and mineral surfaces. Desorption electrospray ionization (DESI) is carried out by directing electrosprayed charged droplets and ions of solvent onto the surface to be analyzed. The impact of the charged particles on the surface produces gaseous ions of material originally present on the surface. The resulting mass spectra are similar to normal ESI mass spectra in that they show mainly singly or multiply charged molecular ions of the analytes. The DESI phenomenon was observed both in the case of conductive and insulator surfaces and for compounds ranging from nonpolar small molecules such as lycopene, the alkaloid coniceine, and small drugs, through polar compounds such as peptides and proteins. Changes in the solution that is sprayed can be used to selectively ionize particular compounds, including those in biological matrices. In vivo analysis is demonstrated.

An awkward feature of mass spectrometry (MS) is that the sample must be introduced into vacuum or into an inaccessible region closely coupled to the vacuum system. Here we describe a simple approach that allows ambient sampling for MS analysis. Electrosprayed (ES) aqueous droplets are directed at a surface of interest in air. The sample can be moved continuously or reoriented in space while MS analysis proceeds. The microdroplets act as projectiles and desorb ions from the surface as a result of electrostatic and pneumatic forces (*1*). The desorbed gas-phase ions are transferred to the distant mass spectrometer via an atmospheric pressure ion-transfer line.

This new method, termed desorption electrospray ionization (DESI), is related to other spray ionization methods, including electrospray ionization (ESI) (*2*, *3*), and to the desorption ionization methods, such as secondary ion mass spectrometry (SIMS) and laser desorption (*4*, *5*). No matrix is needed to perform the experiment—an advantage shared with laser desorption from porous silicon surfaces (*6*)—and the production of multiply charged biological ions is advantageous in extending the mass range, as is the case with ESI.

In its simplest form, the desorption electrospray experiment (Fig. 1) uses an aqueous spray directed at an insulating sample or an analyte deposited on an insulating surface such as polytetrafluoroethylene (PTFE). The desorbed ions are sampled with a commercial ion trap mass spectrometer equipped with an atmospheric interface connected via an extended and preferably flexible ion-transfer line made either of metal or an insulator.

Examination of insulator surfaces is highly unusual in MS; however, the fast nebulizing gas jet in the DESI experiment transports the charged microdroplets and allows them to impact the surface and to carry away analyte molecules. The underlying processes might be related to those occurring in versions of SIMS that use clusters as projectiles (*7*–*11*)

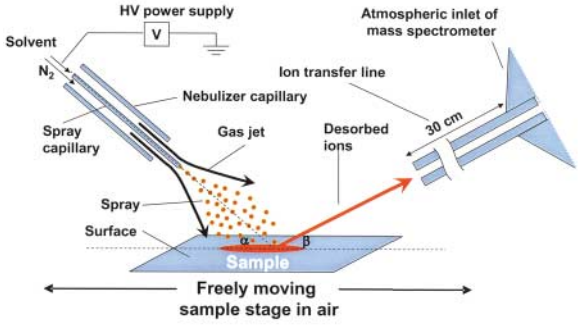


Fig. 1. Schematic of typical DESI experiment. The sample solution was deposited from solution and dried onto a PTFE surface, and methanol-water (1:1 containing 1% acetic acid or 0.1% aqueous acetic acid solution) was sprayed at a flow rate of 3 to 15 μl/min under the influence of a high (4 kV) voltage. The nominal linear velocity of the nebulizing gas was set to 350 m/s.

Department of Chemistry, Purdue University, West Lafayette, IN 47907, USA.

*To whom correspondence should be addressed. E-mail: cooks@purdue.edu

or impact desolvation of droplets on surfaces (*12*) in vacuum environments.

A broad range of analytes have been examined, from simple amino acids through drug molecules, alkaloids, terpenoids, and steroids, to peptides and proteins. The methodology seems to be particularly promising for forensic and public-safety applications, including analysis of dried blood, detection of explosives, and monitoring of chemical warfare agents (*13*), as illustrated by two experiments. In one experiment, the explosive RDX was desorbed from an insulating tanned leather surface, to give the mass spectrum shown in Fig. 2A. In the other experiment, nitrile gloves that were exposed for less than 1 s to dimethyl methylphosphonate vapors (DMMP, a chemical warfare agent simulant), and then washed and dried, gave a mass spectrum (Fig. 2B) that unequivocally indicates the presence of trace levels of DMMP.

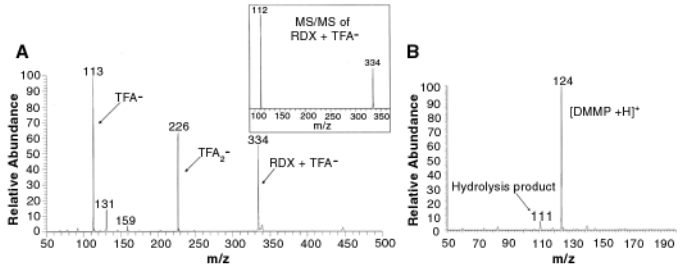


Fig. 2. (**A**) Negative-ion DESI spectrum of RDX (1 ng/mm^2) on leather (porcine) surface using acetonitrile/methanol/trifluoroacetic acid (TFA) 1:1:0.1% as spray solvent. The presence of the explosive is confirmed by tandem MS (inset). (**B**) Positive-ion DESI spectrum of DMMP, a chemical warfare agent simulant, from a nitrile glove surface using the same spray.

New applications of MS might emerge from such simple sampling procedures. In particular, process analysis and other high-throughput experiments are much simplified over standard MS methods. Initial experiments with pharmaceuticals show analysis rates of 20 samples per second (*1*). Optimum experimental conditions are summarized in table S1 (*1*). The ultimate sensitivity of the DESI method has not been determined, but lysozyme present in amounts ranging from 10 to 50 pg could be detected.

A feature of DESI relative to traditional desorption ionization methods, and indeed other MS methods, is the ease with which chemical reagents can be supplied to the site of analysis. This allows the generation of specific reaction products that can be used to confirm analyte identification. Biochemical reactions, including the formation of noncovalent complexes between enzymes and substrates, also serve this purpose. For example, when the lysozyme substrate hexa-(*N*-acetyl) chitohexaose was included in the solution sprayed onto a lysozyme sample, the enzyme-substrate complex was seen at mass-to-charge (*m*/*z*) ratios of 1944 and 2220 (fig. S3) (*1*, *14*–*16*). Another example of a specific chemical reaction used to confirm MS identification is the formation of metal complexes between an analyte on the surface and a metal ion introduced into the spray solution. Uses for this capability include experiments (*1*) in which the chirality of amino acids is measured via diastereomeric complex ion formation and fragmentation (*17*, *18*).

Both MALDI (matrix-assisted laser desorption/ionization) (*19*–*21*) and SIMS (*5*, *22*, *23*) can be used to image biological materials in experiments usually done in vacuum. The exceptions are atmospheric pressure (AP)–MALDI (*21*, *24*) and AP-laser ablation (*25*), but in both of these methods the sample is strictly positioned relative to the rest of the ion source and is inaccessible and not manipulated during the experiment. Working under ambient conditions, DESI can be used for the spatial analysis of native surfaces, such as plant or animal tissues. The potential for this type of application is illustrated by the DESI spectrum of a seed section of poison hemlock (*Conium maculatum*) (Fig. 3A). The peak at *m*/*z* 126 is due to coniceine, which is known to be present in this particular plant species. The

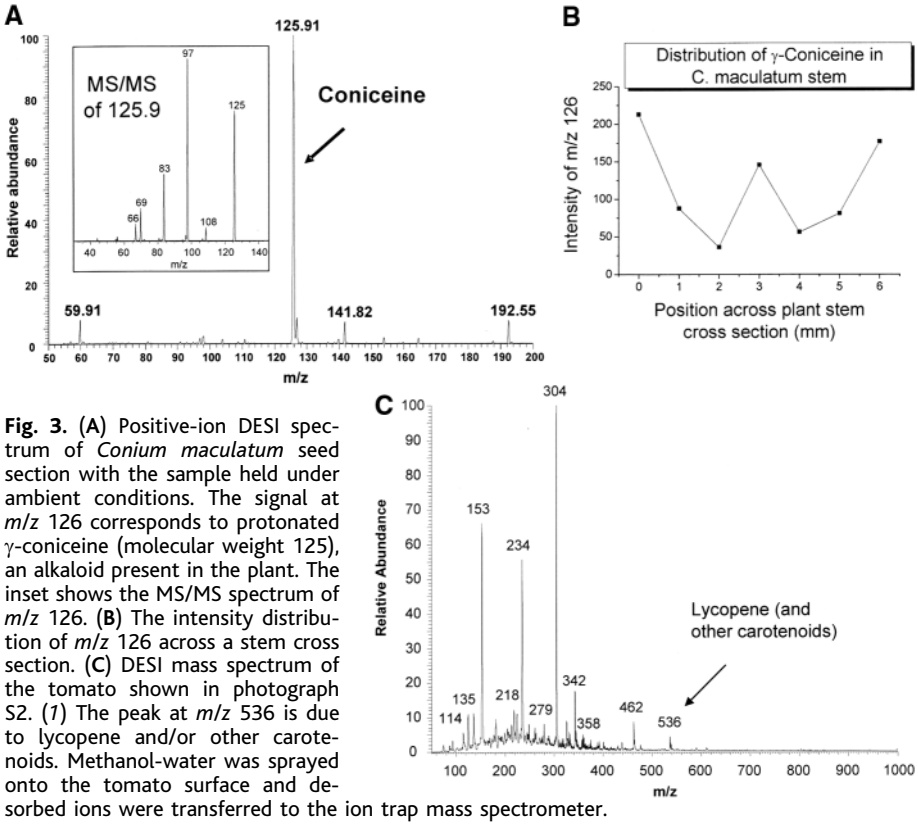


Fig. 3. (**A**) Positive-ion DESI spectrum of *Conium maculatum* seed section with the sample held under ambient conditions. The signal at *m*/*z* 126 corresponds to protonated γ-coniceine (molecular weight 125), an alkaloid present in the plant. The inset shows the MS/MS spectrum of *m*/*z* 126. (**B**) The intensity distribution of *m*/*z* 126 across a stem cross section. (**C**) DESI mass spectrum of the tomato shown in photograph S2. (*1*) The peak at *m*/*z* 536 is due to lycopene and/or other carotenoids. Methanol-water was sprayed onto the tomato surface and desorbed ions were transferred to the ion trap mass spectrometer.

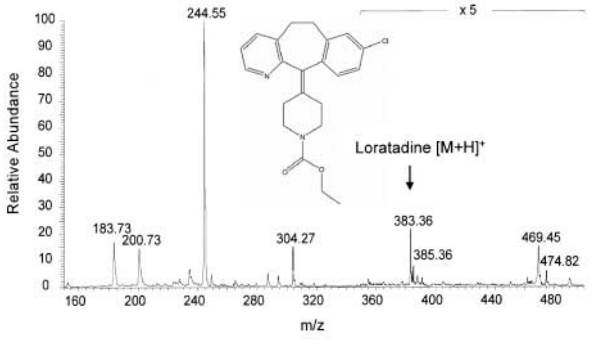


Fig. 4. DESI spectrum recorded by spraying methanol-water onto the finger of a person 50 min after taking 10 mg of the over-the-counter antihistamine Loratadine (*m*/*z* 383/385). Other peaks, e.g., at *m*/*z* 245, are either due to the lab air background or are associated with the detergent that was used for hand washing before the experiment. Note that mass assignments are from the instrument data system and are accurate to just 0.5 *m*/*z* units.

possibility of in situ imaging was demonstrated by scanning the spray spot across a cross section of the plant stem (Fig. 3B). Similarly, the DESI spectrum collected from tomato (*Lycopersicon esculentum*) skin also indicates the localization of characteristic compounds, including lycopene at *m/z* 536 (Fig. 3C).

Quantitative results can be obtained by using appropriate internal standards in experiments where the sample is deposited on a target surface; however, quantification by any method is intrinsically difficult in the analysis of natural surfaces. Sprayed compounds used as internal standards yielded semiquantitative results (relative standard deviation values of ~30%) for spiked plant tissue surfaces.

We have also used DESI for in vivo sampling of living tissue surfaces. An aqueous-alcohol DESI spray was directed onto the finger of a person who had taken 10 mg of the over-the-counter antihistamine Loratadine. About 40 min after taking the tablet, the molecule became detectable directly on the skin or in saliva and its surface concentration remained above the detection limit for another 50 min. A typical spectrum taken directly from skin (photograph S3) is shown in Fig. 4. This example is representative of the novel applications of MS once its vacuum constraints are lifted.

References and Notes

1. Details of the methods and additional experimental results are available on *Science* Online.
2. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science* **246**, 64 (1989).
3. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Mass Spectrom. Rev.* **9**, 37 (1990).
4. A. Benninghoven, F. G. Rudenauer, H. W. Werner, *Secondary Ion Mass Spectrometry: Basic Concepts, Instrumental Aspects, Applications and Trends*, Chemical Analysis, Series of Monographs on Analytical Chemistry and Its Applications (Wiley, New York, 1987), vol. 86.
5. N. Winograd, *Appl. Surf. Sci.* **203**, 13 (2003).
6. J. Wei, J. M. Buriak, G. Siuzdak, *Nature* **399**, 243 (1999).
7. D. Fabris, Z. Wu, C. C. Fenselau, *J. Mass Spectrom.* **30**, 140 (1995).
8. J. F. Mahoney *et al.*, *Rapid Commun. Mass Spectrom.* **5**, 441 (1991).
9. J. F. Mahoney, D. S. Cornett, T. D. Lee, *Rapid Commun. Mass Spectrom.* **8**, 403 (1994).
10. D. S. Cornett, T. D. Lee, J. F. Mahoney, *Rapid Commun. Mass Spectrom.* **8**, 996 (1994).
11. M. G. Blain, E. A. Schweikert, E. F. Dasilveira, *J. Phys.* **50**, 85 (1989).
12. S. A. Aksyonov, P. Williams, *Rapid Commun. Mass Spectrom.* **15**, 2001 (2001).
13. G. L. Gresham *et al.*, *Int. J. Mass Spectrom.* **208**, 135 (2001).
14. The observed complexes show loss of di-*N*-acetylchitobiose under collision-induced dissociation (CID) conditions, clearly indicating the formation of a specific enzyme-substrate complex.
15. F. He, J. Ramirez, C. B. Lebrilla, *Int. J. Mass Spectrom.* **193**, 103 (1999).
16. J. M. Daniel, S. D. Friess, S. Rajagopalan, S. Wendt, R. Zenobi, *Int. J. Mass Spectrom.* **216**, 1 (2002).
17. W. A. Tao, R. G. Cooks, *Anal. Chem.* **75**, 25A (2003).
18. A. Filippi, A. Giardini, S. Piccirillo, M. Speranza, *Int. J. Mass Spectrom.* **198**, 137 (2000).
19. P. Chaurand, S. A. Schwartz, R. M. Caprioli, *Anal. Chem.* **76**, 86A (2004).
20. M. Stoeckli, P. Chaurand, D. E. Hallahan, R. M. Caprioli, *Nature Med.* **7**, 493 (2001).
21. V. V. Laiko, S. C. Moyer, R. J. Cotter, *Anal. Chem.* **72**, 5239 (2000).
22. P. J. Todd, T. G. Schaaff, P. Chaurand, R. M. Caprioli, *J. Mass Spectrom.* **36**, 355 (2001).
23. D. Touboul *et al.*, *Anal. Chem.* **76**, 1550 (2004).
24. V. V. Laiko, M. A. Baldwin, A. L. Burlingame, *Anal. Chem.* **72**, 652 (2000).
25. R. Stockle *et al.*, *Anal. Chem.* **73**, 1399 (2001).

26. This work was supported by Inproteo, Inc. (Indianapolis, IN).

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/471/DC1
Materials and Methods
Table S1
Figs. S1 to S8
Photographs S1 to S3

23 August 2004; accepted 15 September 2004

Vibrational Energy Transfer Across a Reverse Micelle Surfactant Layer

John C. Deàk,[1]* Yoonsoo Pang,[2]* Timothy D. Sechler,[1] Zhaohui Wang,[2] Dana D. Dlott[2]†

In a suspension of reverse micelles, in which the surfactant sodium dioctyl sulfosuccinate (AOT) separates a water nanodroplet from a bulk nonpolar CCl_4 phase, ultrafast vibrational spectroscopy was used to study vibrational energy transfer from the nanodroplet through the AOT interfacial monolayer to the surrounding CCl_4. Most of the vibrational energy from the nanodroplet was transferred to the polar AOT head group within 1.8 picoseconds and then out to the CCl_4 within 10 picoseconds. Vibrational energy pumped directly into the AOT tail resulted in a slower 20- to 40-picosecond transfer of energy to the CCl_4.

The flow of heat between two bulk phases separated by an interfacial monolayer is usually a simple function of the thermal conductivities of the two phases. However, when the heat source is within a few molecular diameters of the interfacial layer, the interfacial thermal conductivity becomes important. The flow of vibrational energy across the interfacial monolayer can depend on how and where the energy is deposited in the system, and different excitations may travel across the interface along different pathways and with different rates.

We present an example of such a situation in a reverse micelle (*1*) system. We used nonlinear vibrational spectroscopy with picosecond time resolution to monitor the flow of energy across surfactant molecules that separate nanodroplets of confined water from a nonpolar liquid phase. We found that vibrational energy deposited in the water could be transferred to the polar surfactant head groups and then to the nonpolar phase in 10 ps; conversely, energy deposited directly in the alkyl surfactant tails was transferred to the nonpolar phase on a longer 20- to 40-ps time scale.

A number of laboratories have studied the ultrafast dynamics of micelle-confined water (*2*). Ultrafast infrared (IR) (*3*, *4*) or THz (*5*) spectroscopies are notable because they do not require the use of extrinsic dopants. Seifert and co-workers (*3*, *4*) looked at sodium dioctyl sulfosuccinate (AOT) surfactant reverse micelles with a water:AOT ratio in the 10 to 55 range. They excited the OH stretch (ν_{OH}) of the confined water (a broad band peaked near 3500 cm^{-1}) with IR pulses, somewhat longer than the ν_{OH} lifetime, that create thermalized confined water. The subsequent cooling of the confined water over hundreds of picoseconds, the nonexponential cooling process, and its dependence on micelle diameter could be explained with ordinary heat conduction theory for a hot water droplet suspended in a colder nonpolar bath (*4*). The success of that explanation indicated that the AOT interface thermal conductivity had little effect on the flow of heat from the water to the nonpolar phase. As a consequence of the rather large water:AOT ratios, the average distance between the heat source and the interfacial layer was large enough that heat diffusion from the interior of the water droplet limited the micelle cooling rate (*4*).

[1]Department of Chemistry, University of Scranton, Scranton, PA 18510, USA. [2]School of Chemical Sciences, University of Illinois at Urbana-Champaign, Box 01-6 Chemical and Life Sciences Laboratory, 600 South Mathews Avenue, Urbana, IL 61801, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: dlott@scs.uiuc.edu

We focus here on energy transfer through the interfacial monolayer using AOT reverse micelles suspended in a nonpolar CCl_4 phase (*6*) with the much lower water:AOT ratio of 2 (Fig. 1). According to simulations and light scattering measurements of AOT/water/CCl_4 reverse micelles (*7*), our samples consisted of micelles about 35 Å in mean diameter with an AOT mean aggregation number of 17 to 18, enclosing water droplets of ~35 molecules. We analyzed IR spectra of the concentration-dependent ν_{OH} line shape, using methods described in (*8*), and found that most of the confined water was interfacial water (*1*, *9*, *10*) bound or trapped near the AOT SO_3^- (sulfonate) head groups. There was little interior or "bulk-like" water, as expected from experiments (*8*) and simulations (*9*) that indicate the interfacial water fraction is 70 to 80%. Our IR pulse pumps this primarily interfacial water on a time scale shorter than the ν_{OH} lifetime. In this case, IR pumping provides a short-duration burst of vibrational energy to the surfactant layer, as opposed to a prolonged heat source emanating from the nanodroplet interior. We also can directly pump the interfacial layer itself by exciting CH stretch (ν_{CH}) excitations of the AOT backbone. Because >90% of the ν_{CH} backbone excitations are in the AOT tail groups, we will henceforth call this "tail excitation." With either water or AOT tail excitation, the "heat source" is a nonthermalized specific vibrational excitation located very near the interface, so that the energy dissipation process cannot be modeled by simple heat conduction.

Instead, the flow of energy across the interface must be understood in the context of specific vibrational energy transfer mechanisms. Thermalization of an OH or CH parent excitation, ultimately leading to a temperature jump (ΔT) of the entire suspension (*11*), involves a multistep process, termed a vibrational cascade, in which parent decay creates successive generations of lower energy daughter vibrations. As each generation is spawned, an amount of energy equal to the specific vibrational energy mismatch is lost to the surroundings. This cascade is generally understood in the context of a three–energy regime picture (*12*). The parents are in the higher energy regime where lifetimes are short (say 1 to 5 ps) because of a high density of states. Daughters in the intermediate regime (typically 500 to 1700 cm^{-1}) have the longest lifetimes (say 5 to 20 ps) because of the lower density of states. Daughters in the lower energy regime again have short lifetimes, being large-amplitude molecular deformations with strong anharmonic coupling. The CCl_4 is heated by the energy from vibrational mismatches and from large-amplitude deformations. This energy is rapidly equilibrated among CCl_4 vibrations, so that CCl_4 vibrations function as a molecular thermometer with a <10-ps response time (*11*, *13*, *14*).

Our experimental technique (*15*) combined IR pumping with incoherent anti-Stokes Raman probing (*16*). The anti-Stokes Raman signal is sensitive only to vibrational excited states. The signal strength was proportional to the Raman cross section times the vibrational occupation number, and the cross sections are available from the Stokes Raman spectrum in Fig. 1. We simultaneously monitored the decay of parent excitations and the rise and subsequent decay of successive generations of daughters. In the reverse micelle system, many of these excitations can be associated with specific features of micelle structure (*17*, *18*): the OH stretch (ν_{OH}) of the confined water; the sulfonate stretches (symmetric ν^s_{SO3} and asymmetric ν^a_{SO3}) and the carbonyl stretch (ν_{CO}) of the AOT head group; the CH stretches (ν_{CH}), CH bends (δ_{CH}), and CC stretches (ν_{CC}) of the AOT alkyl tail; and the C-Cl stretches of CCl_4. Watching vibrational energy redistribute with high time resolution thus provides a degree of spatial resolution that is a function of the degree of spatial localization of each vibrational excitation and foreknowledge of the structure (*19*). In micelles, we can clearly distinguish whether an excitation is instantaneously in the nanodroplet, the AOT head or tail, or the surrounding CCl_4.

Figure 2 shows a series of anti-Stokes Raman spectra acquired with increasing delay after 3500 cm^{-1} ν_{OH} pumping of micelle-confined water. About 5% of the water molecules are excited (*20*, *21*), one or two per micelle on average. The IR pulse also excited the carbonyl stretching overtone $2\nu_{CO}$. The amount of $2\nu_{CO}$ excitation was small compared to the amount of ν_{OH} excitation. Tuning the ν_{OH} pump pulse over ~200 cm^{-1} varied the amount of $2\nu_{CO}$ at constant ν_{OH} but had little effect on the vibrational energy transfer. The case of AOT ν_{CH} alkyl tail excitation (*14*, *22*–*24*) is shown in the series of anti-Stokes Raman spectra in Fig. 3.

The traces in Fig. 4 reveal the dynamics extracted from the spectra in Figs. 2 and 3. We focused first on water pumping. We have conducted detailed studies of bulk water

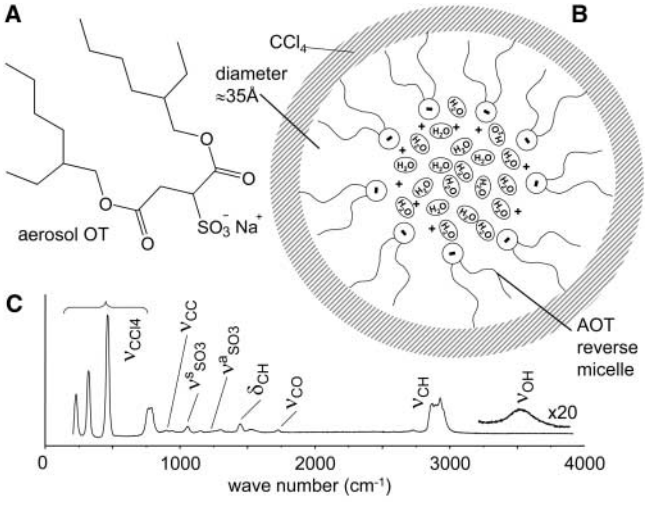


Fig. 1. (**A**) Structure of the AOT surfactant. (**B**) Schematic of a reverse micelle with a water:AOT concentration of 2. (**C**) Assignments of the Raman spectrum of reverse micelles in CCl_4. The ν_{OH} region was multiplied by a factor of 20 for clarity.

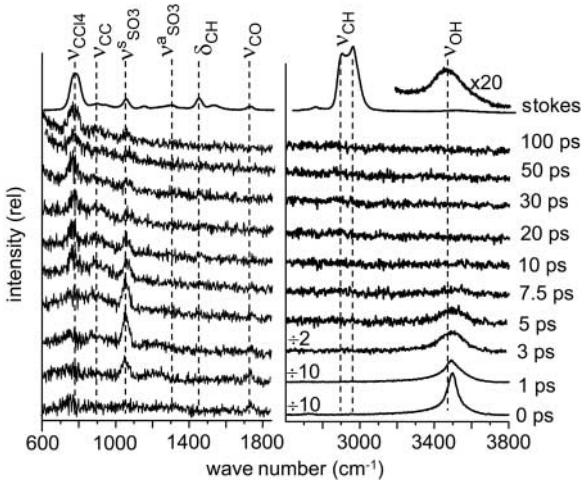


Fig. 2. Top trace: Stokes Raman spectrum of AOT micelle suspension. Lower traces: A time series of transient anti-Stokes spectra after OH stretch pumping of confined water at 3500 cm^{-1}. rel, relative units.

(*20*, *21*, *25*), which evidences complicated wavelength-dependent dynamics and a coherent artifact due to nonlinear light scattering (NLS) that appears at the IR pump pulse wave number. On the basis of those studies, we extracted the NLS contribution shown in Fig. 4, a Gaussian function with half-width at half-maximum (HWHM) of 0.55 ps that represents the laser apparatus instantaneous time response (*21*). The time dependence of the ν_{OH} spectrum was used to determine an effective exponential decay time constant T_1 = 1.8 (±0.1) ps, more than twice the 0.75-ps lifetime of bulk water pumped at 3500 cm^{-1} (*21*). This lifetime increase is similar to what has been seen recently in the ν_{OD} of micelle-confined interfacial HOD (*26*) and in the ν_{OH} of HOD molecules in the aqueous solvation shell of the I^- anion (*27*). It likely has a similar cause: reduced anharmonic coupling to the bath resulting from anion-induced weakening of the hydrogen bonding (*1*, *8*–*10*, *28*). As energy leaves the ν_{OH} vibration, a substantial portion passes into the ν^s_{SO3} and ν^a_{SO3} head group stretches: Fig. 4A shows ν^s_{SO3} excitation rising in 1.8 ps as ν_{OH} decays and decaying with a T_1 of 6 (±1) ps. The sulfonate head group excitations were the predominant first-generation daughters; little AOT tail excitation was observed.

Figure 4B shows the dynamics after ν_{CH} excitation of the AOT tails. The parent ν_{CH} excitations (*22*) decayed with a T_1 of 2.3 (±0.1) ps. The first-generation daughters include δ_{CH} and ν_{CC} tail excitations plus just a small amount of head group stretching. Fitting the data with a model of exponential buildup and decay gives effective lifetimes of 4.4 ps for δ_{CH} and 5 to 10 ps for ν_{CC}. After ν_{CH} pumping, ν^s_{SO3} and ν_{CC} excitations were seen at longer delay times than were observed with ν_{OH} pumping. These populations result not because the ν^s_{SO3} and ν_{CC} lifetimes change, but because the ν_{CH} decay process spawns the daughter vibrations over a longer time period (see below). Pumping the AOT ν_{CH} stretch leads to less head group excitation but much more AOT tail excitation than we observed after pumping the water.

Figure 4C details the rise of the 780 cm^{-1} signal from the CCl_4 molecular thermometer. The bulk temperature jump ΔT depends on the pump pulse energy and absorption coefficient (*14*). After thermalization, ΔT was 5 K for water pumping and 16 K for AOT pumping, and to facilitate comparison of the time dependences we normalized the data in Fig. 4C to the same final temperature. With water pumping, most of the CCl_4 temperature rise occurred in the first 10 ps. With AOT tail pumping, the CCl_4 temperature did not begin to rise until ~20 ps, with equilibrium attained within ~40 ps.

If energy transfer from the reverse micelle to the surrounding CCl_4 operated through simple heat conduction, pumping the AOT tail would result in faster CCl_4 heating than would confined water pumping. However, CCl_4 actually heats up faster after water pumping. In order to explain this result, we have to consider the specific vibrational relaxation pathways in each case.

Water pumping delivers a 1.8-ps burst of ~3500 cm^{-1} energy to AOT. Figure 4C indicates that during this burst about one-fifth of the energy is dissipated to CCl_4, with the remaining four-fifths going to the sulfonate head group and to a lesser extent the AOT tail. The route from sulfonate to CCl_4 is seen to be rapid and efficient because all subsequent CCl_4 heating appears on the same 5- to 10-ps time scale as sulfonate stretch decay. Based on prior studies of NO_2 (*29*), we believe that sulfonate stretches excite sulfonate rocking and scissoring modes that transmit energy efficiently to tail wagging and CCl_4. AOT pumping generates mainly long-lived δ_{CH} and ν_{CC} that, being near the top of the intermediate regime, can produce at least one subsequent generation of long-lived daughters (*14*, *23*, *24*) before all energy is converted into low-frequency molecular deformations. The result is CCl_4 heating primarily during the 20- to 40-ps time period.

The fact that micelle cooling from a hot water nanodroplet is faster than micelle cooling with AOT tail excitation is an illustration that energy transfer between two liquid phases separated by a surfactant layer must be analyzed in terms of specific vibrational couplings, rather than ordinary heat conduction.

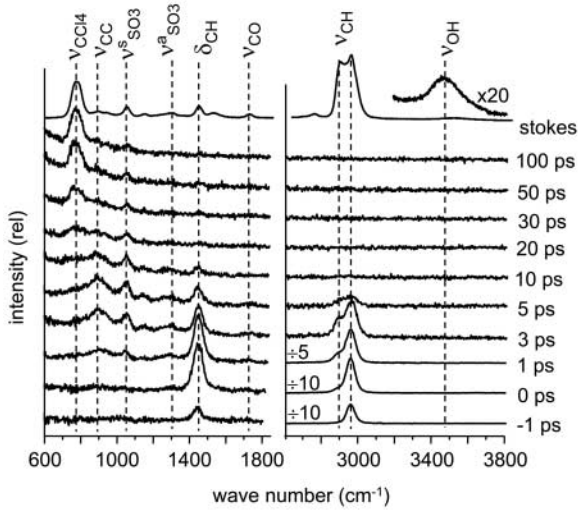


Fig. 3. Top traces: Stokes Raman spectrum of AOT micelle suspension. Lower traces: A series of transient anti-Stokes spectra (as in Fig. 2) acquired at successively longer delay times after CH stretch pumping of AOT at 2950 cm^{-1}.

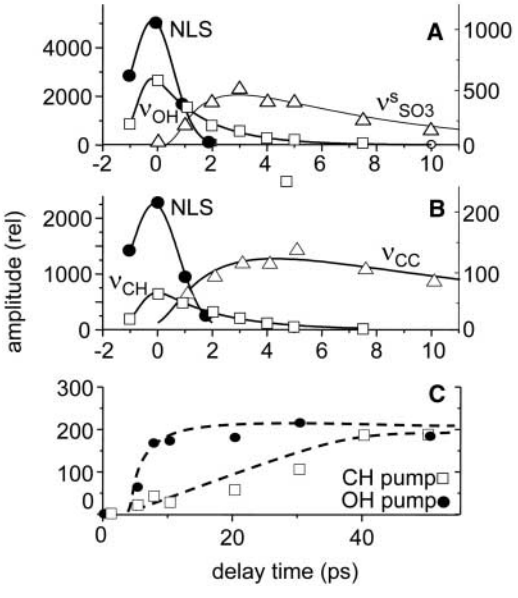


Fig. 4. Dynamics extracted from Figs. 2 and 3. The NLS signal, a coherent artifact that tracks the apparatus-limited time response, is fit to a Gaussian with a HWHM of 0.55 ps. The other smooth curves are convolutions of this Gaussian with an exponential buildup and decay yielding the time constants given in the text. (**A**) Decay of the parent OH stretch ν_{OH} of water (left axis) and subsequent rise and decay of the symmetric stretch ν^s_{SO3} of the AOT head group (right axis). (**B**) Decay of the parent CH stretch ν_{CH} of AOT (left axis) and subsequent rise and decay of the AOT tail stretch ν_{CC} (right axis). (**C**) Rise of the CCl_4 molecular thermometer after OH pumping of confined water and CH pumping of AOT. Both CCl_4 signals were normalized to the same final temperature. The lines through the data are guides to the eye.

References and Notes

1. T. K. De, A. Maitra, *Adv. Colloid Interface Sci.* **59**, 95 (1995).
2. N. E. Levinger, *Science* **298**, 1722 (2002).
3. T. Patzlaff, M. Janich, G. Seifert, H. Graener, *Chem. Phys.* **2000**, 381 (2000).

4. G. Seifert, T. Patzlaff, H. Graener, *Phys. Rev. Lett.* **88**, 147402 (2002).
5. J. E. Boyd, A. Briskman, C. M. Sayes, D. Mittleman, V. Colvin, *J. Phys. Chem. B* **106**, 6346 (2002).
6. E. Sein, J. R. Lalanne, J. Buchert, S. Kielich, *J. Colloid Interface Sci.* **72**, 363 (1979).
7. B. Derecskei, A. Derecskei-Kovacs, Z. A. Schelly, *Langmuir* **15**, 1981 (1999).
8. G. Onori, A. Santucci, *J. Phys. Chem.* **97**, 5430 (1993).
9. J. Faeder, B. M. Ladanyi, *J. Phys. Chem. B* **104**, 1033 (2000).
10. D. Brown, J. H. R. Clarke, *J. Phys. Chem.* **92**, 2881 (1988).
11. L. K. Iwaki, D. D. Dlott, *J. Phys. Chem. A* **104**, 9101 (2000).
12. A. Nitzan, J. Jortner, *Mol. Phys.* **25**, 713 (1973).
13. P. B. Graham, K. J. M. Matus, R. M. Stratt, *J. Chem. Phys.* **121**, 5348 (2004).
14. L. K. Iwaki, J. C. Deàk, S. T. Rhea, D. D. Dlott, in *Ultrafast Infrared and Raman Spectroscopy*, M. D. Fayer, Ed. (Marcel Dekker, New York, 2000), pp. 541–592.
15. Materials and methods are available as supporting material on *Science* Online.
16. A. Laubereau, W. Kaiser, *Rev. Mod. Phys.* **50**, 607 (1978).
17. P. D. Moran, G. A. Bowmaker, R. P. Cooney, *Langmuir* **11**, 738 (1995).
18. P. D. Moran, G. A. Bowmaker, R. P. Cooney, J. R. Bartlett, J. L. Woolfrey, *J. Mater. Chem.* **5**, 295 (1995).
19. Z. Wang, A. Pakoulev, D. D. Dlott, *Science* **296**, 2201 (2002).
20. J. C. Deàk, L. K. Iwaki, D. D. Dlott, *J. Phys. Chem.* **104**, 4866 (2000).
21. A. Pakoulev, Z. Wang, Y. Pang, D. D. Dlott, *Chem. Phys. Lett.* **380**, 404 (2003).
22. The 2950 cm^{-1} IR pulse was tuned to the higher energy part of the ν_{CH} absorption (*15*), dominated by ν^a_{CH} asymmetric stretches. Only 3 of the 34 H atoms of the AOT backbone are in the head group, so the excitations were predominantly distributed throughout the AOT tails. Because of a Fermi resonance, the parent was an admixture of ν^a_{CH} and $2\delta^a_{CH}$ bend overtones. The $2\delta^a_{CH}$ character is seen in the $v = 2 \rightarrow 1$ transition (where v is the vibrational quantum number), an unresolved tail ~30 cm^{-1} to the red of the fundamental transition at 1460 cm^{-1}. Energy redistribution among ν^a_{CH} and ν^s_{CH} modes is also seen in the delayed appearance of ν^s_{CH} signal at ~2900 cm^{-1}. These processes are similar to what has been seen in other alkane systems (*14*, *22*, *23*).
23. J. C. Deàk, L. K. Iwaki, D. D. Dlott, *J. Phys. Chem.* **102**, 8193 (1998).
24. J. C. Deàk, L. K. Iwaki, S. T. Rhea, D. D. Dlott, *J. Raman Spectrosc.* **31**, 263 (2000).
25. Z. Wang, A. Pakoulev, Y. Pang, D. D. Dlott, *Chem. Phys. Lett.* **378**, 281 (2003).
26. H.-S. Tan, I. R. Piletic, R. E. Riter, N. E. Levinger, M. D. Fayer, unpublished data.
27. M. F. Kropman, H. J. Bakker, *Chem. Phys. Lett.* **370**, 741 (2003).
28. D. S. Venables, K. Huang, C. A. Schmuttenmaer, *J. Phys. Chem. B* **105**, 9132 (2001).
29. J. C. Deàk, L. K. Iwaki, D. D. Dlott, *J. Phys. Chem. A* **103**, 971 (1999).

30. This material is based on work supported by the U.S. Department of Energy, Division of Materials Sciences under award no. DEFG02-91ER45439, through the Frederick Seitz Materials Research Laboratory at the University of Illinois at Urbana-Champaign; by the National Science Foundation under award no. DMR-0096466, and by the Air Force Office of Scientific Research under award no. F49620-03-1-0032. J.C.D. acknowledges support from the Office of Research Services at the University of Scranton.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1102074/DC1
Materials and Methods
Figs. S1 and S2
References and Notes

29 June 2004; accepted 13 September 2004
Published online 23 September 2004;
10.1126/science.1102074
Include this information when citing this paper.

A Network of Superconducting Gravimeters Detects Submicrogal Coseismic Gravity Changes

Yuichi Imanishi,[1]* Tadahiro Sato,[2] Toshihiro Higashi,[3] Wenke Sun,[4] Shuhei Okubo[4]

With high-resolution continuous gravity recordings from a regional network of superconducting gravimeters, we have detected permanent changes in gravity acceleration associated with a recent large earthquake. Detected changes in gravity acceleration are smaller than 10^{-8} meters seconds^{-2} (1 microGalileo, about 10^{-9} times the surface gravity acceleration) and agree with theoretical values calculated from a dislocation model. Superconducting gravimetry can contribute to the studies of secular gravity changes associated with tectonic processes.

An earthquake is a discontinuous displacement along a fault plane within Earth (*1*) and gives rise to a change in Earth's mass distribution as well as the radiation of seismic waves. This in turn results in a change in the gravity potential field of Earth. Viewed from a point fixed with respect to Earth, this should be observed as a small change in local gravity acceleration through two effects: apparent addition or subtraction of Earth's mass and a change in the distance to the center of Earth. Such coseismic gravity changes have been observed at the 1964 M_w (moment magnitude) = 9.2 Alaska earthquake (*2*), the 1998 M = 6.1 earthquake near Mount Iwate-san (*3*), and the 2000 dike-intrusion event at the Miyake-jima volcano (*4*). A theory (*5*–*8*) has been developed to calculate coseismic gravity changes for arbitrary fault geometries on the basis of the dislocation model of the earthquake source and has been shown to reproduce the observed values well. Because coseismic gravity changes depend on fault parameters, especially dislocation vectors (*5*), their observations at multiple stations, if combined with other techniques such as global positioning system and strainmeters, should provide constraint on the nature of dislocations in the earthquake source region. However, the previous observations must be regarded as rather exceptional; they mostly have their basis in gravity values obtained before and after the events over long time intervals and therefore are subject to ambiguities due to temporal instability of the gravimeters. In addition, limited precision of the gravimeters has confined observations to the near field (typical epicentral distances are less than 1°) where relatively large gravity changes, on the order of 10^{-7} m s^{-2}, are expected. Systematic detection of coseismic gravity changes on more common earthquakes in intermediate to far fields by means of continuous gravity observations has long been desired.

Here, we report the observation of earthquake-induced gravity changes with magnitudes less than 10^{-8} m s^{-2} (1 μGal) from continuous gravity recordings. This has been made possible by the high sensitivity and stability of superconducting gravimeters (SGs) (*9*). The SG provides an ideal tool for precisely monitoring tidal and nontidal temporal gravity changes, including free oscillations of the Earth (*10*), environmental effects (*11*), and gravity waves in the fluid core of Earth (*12*). As part of the Global Geodynamics Project network (*13*), gravity observations with SGs are made at stations Esashi (*14*), Matsushiro (*15*), and Kyoto (*16*) in Japan (Fig. 1). Incidentally, these stations are aligned roughly along a straight line, thus forming a domestic array of SGs.

On 25 September 2003 [universal time coordinated (UTC)], a large earthquake (M_w = 8.0) (*17*), the Tokachi-oki event, occurred close to Hokkaido Island. The epicenter (Fig. 1) is almost in line with the SG array. Epicentral distances to the SG stations are 3.4°, 6.9°, and 9.4° for Esashi, Matsushiro, and Kyoto, respectively. The 2003 Tokachi-oki earthquake was a thrust fault event occurring along the plate interface where the Pacific Plate is subducting beneath the North American Plate. A thrust earthquake dis-

[1]Ocean Research Institute, University of Tokyo, 1-15-1, Minamidai, Nakano, Tokyo 164-8639, Japan. [2]National Astronomical Observatory of Japan, 2-12, Hoshigaoka, Mizusawa, Iwate 023-0861, Japan. [3]Department of Geophysics, Graduate School of Science, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan. [4]Earthquake Research Institute, University of Tokyo, 1-1-1, Yayoi, Bunkyo, Tokyo 113-0032, Japan.

*To whom correspondence should be addressed. E-mail: imanishi@ori.u-tokyo.ac.jp

places adjacent blocks vertically, producing permanent changes in the gravity field around the epicentral region. Moreover, seismic coupling for the asperity associated with this earthquake is reported to have been about 100% (*17*), implying that there were no substantial slow-slip events before and after the main shock. In short, this earthquake provided an ideal opportunity to investigate how the SG network recorded coseismic gravity changes.

Dislocation model theory (*5*–*8*) can incorporate different degrees of elaboration in models of Earth structure and earthquake source to predict gravity changes. Even the simplest model, which treats Earth as a homogeneous half-space (*5*), has sufficient precision in the near field to be compared with observations. However, it is not applicable to the 2003 Tokachi-oki event, because epicentral distances from the SG stations are greater than the characteristic spatial scale of the earthquake source. A spherical and homogeneous Earth model gives similar results in the near field, but the effect of radial layering of the Earth is not negligible in the far field (*7*). Therefore, we must apply a spherical and layered Earth model, specifically symmetric model 1066A (*18*), to the 2003 Tokachi-oki event. Calculations for the spherically symmetric model have been made computer-efficient through the reciprocity theorem (*6*) relating coseismic and tidal or load deformations of Earth. As for the earthquake source, a point source model (*7*) rather than a finite fault plane model (*8*) suffices for the present study, again because epicentral distances are sufficiently long. We used published fault parameters (*17*) (Table 1) in the calculations.

Gravity changes recorded by a gravimeter consist of two terms: a contribution of internal mass redistribution and the effect of vertical displacement of Earth's surface. Gravity observation alone cannot differentiate between the two. The latter is given by $-\beta\Delta h$, where β is the vertical gradient of gravity acceleration and Δh is vertical displacement (positive upward). In theoretical calculations, one normally uses the free-air gravity gradient at Earth's surface (3.09×10^{-6} s^{-2}) for β. Because the Matsushiro station is underground, however, use of the surface gradient would overestimate the effect of vertical displacement on gravity. Therefore, we have used $\beta = 2.28 \times 10^{-6}$ s^{-2} (*19*) for this station. It should be noted that the redistribution and displacement terms (Table 2) contributing to total gravity changes are not simple functions of epicentral distance; vertical displacement dominates at Esashi, whereas it is subordinate at Matsushiro. At Kyoto, the two terms have similar magnitudes and different signs so that they almost cancel. The complex relations of these terms to station coordinates originate from the fault geometry. The total of the two terms at the three stations decreases monotonically with epicentral distance. Considering the high sensitivity of SGs, the predicted total gravity change is well above the detection limit of the instrument at Esashi and in the detection range at Matsushiro and Kyoto.

All three stations provided data for analysis of the 2003 Tokachi-oki earthquake. To determine changes in gravity associated with the event, we removed known signals from the data. Many phenomena cause temporal changes in surface gravity. The largest effect is lunar and solar tidal deformation of the solid Earth. Oceanic tides have a minor contribution to the observed gravity through loading. The combined signals of body and oceanic tides are accurately known for each station from years of gravity observations and are readily removed from the data. The atmosphere affects gravity through Newtonian attraction by the atmospheric mass and deformation of land due to loading. Although this effect could, in principle, be fully understood only by taking into account the global distribution of atmospheric pressure (*20*), it is roughly proportional to the change in local atmospheric pressure, especially for short periods with an admittance around -3×10^{-9} m s^{-2} hPa^{-1} (*21*). In this study, we applied a local atmospheric correction with admittance adjusted for each station. Lastly, the effects of polar motion (long-periodic motion of the instantaneous rotation axis of Earth) and instrumental drift of the gravimeter must be removed from the gravity data. The former can be calculated from Earth rotation parameters provided by the International Earth Rotation and Reference Systems Services (*22*), and the latter can be modeled as a linear function of time.

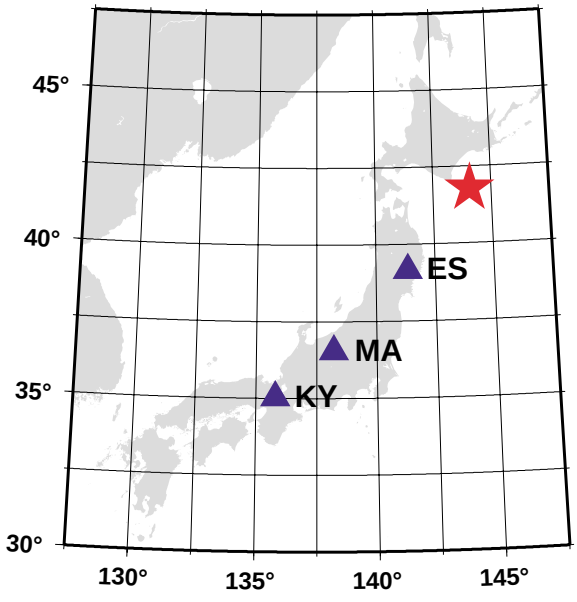


Fig. 1. Epicenter (asterisk) of the 2003 Tokachi-oki earthquake and SG station locations (triangles). ES, Esashi; MA, Matsushiro; KY, Kyoto.

Table 1. Fault parameters of the 2003 Tokachi-oki earthquake used in theoretical calculations.

Parameter	
Latitude	41.780° N
Longitude	144.079° E
Strike	230°
Dip	20°
Rake	109°
Depth	32 km
Fault area	90 km by 70 km
Dislocation	2.6 m

Table 2. Calculated and observed gravity changes at the SG stations for the 2003 Tokachi-oki earthquake. 1 μGal = 10^{-8} m s^{-2}. Uncertainties of the observed values are on the order of 0.1 μGal.

Station	Latitude	Longitude	Epicentral distance	Calculated value (μGal)			Observed value (μGal)
				Redistribution	Displacement	Total	
Esashi	39.151°	141.332°	3.4°	+0.134	+0.485	+0.619	+0.58
Matsushiro	36.543°	138.207°	6.9°	+0.171	−0.048	+0.123	+0.10
Kyoto	35.028°	135.786°	9.4°	−0.239	+0.296	+0.057	+0.07

Both effects are relatively long period, and their correction makes little difference to the final results. After removing the known components, residual gravity should consist of unmodeled (mainly environmental) signals originating from other parts of Earth as well as possible step changes associated with the earthquake.

Normally, processing of SG data to obtain residual gravity is straightforward. In this analysis, however, several factors make interpretation of the Matsushiro and Kyoto data complicated. They include the effect of rainfall and the tilt of the gravimeter [Supporting Online Material (SOM) Text]. No such complications affect data from the Esashi station, which is the closest of the three stations to the epicenter and therefore the most important for this study.

Final estimates of the residual gravity series for the three stations, low-pass filtered and decimated to provide one sample per minute, are not continuous (Fig. 2). The earthquake occurred at 19:50:08 on 25 September 2003 (UTC), and the seismic waves arrived at all three stations soon thereafter. Large accelerations due to high-frequency seismic waves tend to saturate the gravimeters, rendering the gravity data at the time of the earthquake and shortly thereafter unusable. Even if the gravity signal is not saturated, large horizontal accelerations can degrade the performance of the gravimeter (*23*). For this reason, we discarded the first 24 hours of the residual series after the earthquake. Also, we filtered out spikes due to aftershocks. The Kyoto data after 10:00:00 on 28 September 2003 (UTC) were not used for the analysis, because corresponding tiltmeter data are unavailable.

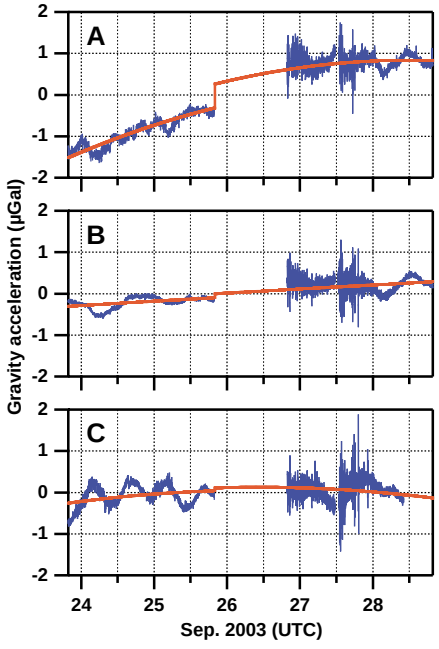


Fig. 2. Residual gravity data (blue) recorded by SGs for the 2003 Tokachi-oki earthquake. (**A**) Esashi, (**B**) Matsushiro, and (**C**) Kyoto. Red curves are best-fit functions given by Eq. 1.

Comparing the trends of the residual series before and after the earthquake, an offset in gravity is evident in the data from the Esashi station, even if taking into account an apparent change in the trend in the 24 hours after the earthquake (SOM Text). The offset is positive, meaning an increase in gravity. Similar offsets, if any, may exist in data from the other two stations but with smaller magnitudes. To quantify this observation, we fit a quadratic time function to the residual series, with an assumed offset in the constant term at the source time:

$$g(t) = a + b(t - t_0) + c(t - t_0) + d\Theta(t - t_0) \quad (1)$$

where $g(t)$ is the residual gravity at time t, t_0 is the source time, and Θ is a step function. We adopt a quadratic function because the residual series for Esashi and perhaps Kyoto show different trends before and after the earthquake, making it difficult to fit simple linear functions to them. A weighted least squares fit of the above function to five days of residual gravity data yields estimates of the gravity offset (d) as 0.575 ± 0.007, 0.096 ± 0.005, and 0.070 ± 0.009 μGal for Esashi, Matsushiro, and Kyoto, respectively. Use of the μGal unit emphasizes that these values are less than 1 μGal, a practical precision limit of typical gravimeters both absolute and relative. The formal uncertainties for these estimates of gravity offset are on the order of 0.01 μGal. It would not be appropriate, however, to quote them simply as estimation errors, because there are correlations among model parameters, especially between b and d. Moreover, an estimate of the offset depends to some degree on the choice of the fit function. For example, using a cubic time function instead of a quadratic one gives 0.443 ± 0.012, 0.166 ± 0.006, and 0.639 ± 0.011 μGal for the three stations. Therefore, uncertainties of 0.1 μGal seem justified. Taking these uncertainties into account, we conclude that positive gravity changes have been detected at Esashi, less definitely at Matsushiro, and marginally at Kyoto.

Agreement between observed and calculated gravity changes is remarkable (Table 2). At Esashi, where the best signal-to-noise ratio is expected for this earthquake, both sign and magnitude agree. The other two stations have also yielded relatively good results in spite of greater epicentral distances and some ambiguities in interpretation of the residual data. Whereas our results help validate the theory for calculating coseismic gravity changes, they also demonstrate that SGs are capable of detecting step changes in gravity as small as, or less than, 1 μGal caused by earthquakes.

Since their invention (*24*), SGs have been applied mostly to studies of global phenomena involving more or less periodic changes of gravity and rarely to aperiodic, which are commonly step changes associated with tectonic processes. Our success in detecting coseismic changes demonstrates the potential of SGs as a high-precision probe of secular gravity changes. Therefore, SG data may be used also as ground truth for satellite gravity missions (*25*), e.g., the Challenging Minisatellite Payload (CHAMP) (*26*) and the Gravity Recovery and Climate Experiment (GRACE) (*27*).

References and Notes

1. K. Aki, P. Richards, *Quantitative Seismology* (University Science, Sausalito, CA, ed. 2, 2002).
2. D. F. Barnes, *J. Geophys. Res.* **71**, 451 (1966).
3. Y. Tanaka *et al.*, *Geophys. Res. Lett.* **28**, 152979 (2001).
4. M. Furuya *et al.*, *J. Geophys. Res.* **108**, 10.1029/2002JB001989 (2003).
5. S. Okubo, *J. Geophys. Res.* **97**, 7137 (1992).
6. S. Okubo, *Geophys. J. Int.* **115**, 921 (1993).
7. W. Sun, S. Okubo, *Geophys. J. Int.* **114**, 569 (1993).
8. W. Sun, S. Okubo, *Geophys. J. Int.* **132**, 79 (1998).
9. J. M. Goodkind, *Rev. Sci. Instrum.* **70**, 4131 (1999).
10. S. Rosat, J. Hinderer, L. Rivera, *Geophys. Res. Lett.* **30**, 10.1029/2003GL018304 (2003).
11. H. Virtanen, *J. Geod. Soc. Japan* **47**, 328 (2001).
12. D. Smylie, *Science* **255**, 1678 (1992).
13. D. Crossley *et al.*, *Eos* **80**, 121125 (1999).
14. T. Sato, Y. Tamura, in *Non-Tidal Gravity Changes: Intercomparison Between Absolute and Superconducting Gravimeters*, C. Poitevin, Ed. (Cahiers du Centre Européen de Géodynamique et de Séismologie, Luxembourg, 1991), vol. 3, p. 109.
15. Y. Imanishi, in *High Precision Gravity Measurements with Application to Geodynamics and Second GGP Workshop*, B. Ducarme, J. Barthélemy, Eds. (Cahiers du Centre Européen de Géodynamique et de Séismologie, Luxembourg, 2000) vol. 17, p. 97.
16. T. Higashi, *Mem. Fac. Sci. Kyoto Univ. Ser. Phys. Astrophys. Geophys. Chem.* **39**, 313 (1996).
17. Y. Yamanaka, M. Kikuchi, *Earth Planets Space* **e21** (2003).
18. F. Gilbert, A. M. Dziewonski, *Phil. Trans. R. Soc. London Ser. A* **278**, 187 (1975).
19. Geographical Survey Institute, Japan, unpublished data.
20. J. P. Boy, P. Gegout, J. Hinderer, *Geophys. J. Int.* **149**, 534 (2002).
21. R. J. Warburton, J. M. Goodkind, *Geophys. J. R. Astron. Soc.* **48**, 281 (1977).
22. More information is available online at www.iers.org/.
23. Y. Imanishi, in preparation.
24. W. A. Prothero, J. M. Goodkind, *Rev. Sci. Instrum.* **39**, 1257 (1968).
25. D. Crossley, J. Hinderer, M. Llubes, N. Florsch, *Adv. Geosci.* **1**, 65 (2003).
26. More information is available online at http://op.gfz-potsdam.de/champ/index_CHAMP.html.
27. More information is available online at www.csr.utexas.edu/grace/.

28. The authors would like to thank S. Takemoto, Y. Fukuda, Y. Tamura, Y. Yamanaka, S. Iwano, S. Ogasawara, and the staff of Matsushiro Seismological Observatory, Japan Meteorological Agency, for their valuable support and discussions and M. F. Coffin and W. Zürn for reading the manuscript carefully.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/476/DC1
SOM Text
Figs. S1 to S4

23 June 2004; accepted 15 September 2004

Local Endemism Within the Western Ghats–Sri Lanka Biodiversity Hotspot

Franky Bossuyt,[1,2*†] Madhava Meegaskumbura,[3,4*] Natalie Beenaerts,[1*] David J. Gower,[5] Rohan Pethiyagoda,[4] Kim Roelants,[1] An Mannaert,[1] Mark Wilkinson,[5] Mohomed M. Bahir,[4] Kelum Manamendra-Arachchi,[4] Peter K. L. Ng,[6] Christopher J. Schneider,[3] Oommen V. Oommen,[7] Michel C. Milinkovitch[2]

The apparent biotic affinities between the mainland and the island in the Western Ghats–Sri Lanka biodiversity hotspot have been interpreted as the result of frequent migrations during recent periods of low sea level. We show, using molecular phylogenies of two invertebrate and four vertebrate groups, that biotic interchange between these areas has been much more limited than hitherto assumed. Despite several extended periods of land connection during the past 500,000 years, Sri Lanka has maintained a fauna that is largely distinct from that of the Indian mainland. Future conservation programs for the subcontinent should take into account such patterns of local endemism at the finest scale at which they may occur.

Island biota typically are closely related to the source of colonists when both areas have been in regular contact (*1–3*). The level of endemism on continental islands is therefore expected to reflect the number and duration of ocean-level lowstands that allowed exchange with the mainland (*4*). Sri Lanka is a relatively large island (~66,000 km^2) in the Indian Ocean and is part of the same shallow continental shelf as India (*5*). During the Pleistocene ice ages, Sri Lanka was intermittently connected to mainland India (*6*), until sea level rise created the present disruption ~10,000 years ago (*7*) (Fig. 1). Classical comparisons of faunal elements from both sides of the Palk Strait indicate a high degree of morphological similarity in several groups, suggesting abundant, recent biotic interchange with southern India (*8–12*). Similar observations prompted Wallace (*13*) more than a century ago to recognize a Ceylonese (or Lankan) biogeographic region, associating Sri Lanka with the southernmost part of the Western Ghats, a hill range along the west coast of India (Fig. 1A). Today, both areas are united in the Western Ghats–Sri Lanka biodiversity hotspot, because they are construed as forming "a community of species that fits together as a biogeographic unit" (*14*).

Here we explore the evolutionary relationships between the subcontinent's island and mainland fauna in two invertebrate and four vertebrate groups. The selected taxa are freshwater crabs (Parathelphusidae and Gecarcinucidae), freshwater shrimps (*Caridina*, Atyidae), tree frogs (*Philautus*, Rhacophorinae, Ranidae), caecilian amphibians (Ichthyophiidae and Uraeotyphlidae), shieldtail snakes (Uropeltidae), and freshwater fishes (*Puntius*, Cyprinidae). These animals occupy a diverse range of habitats (terrestrial, subterranean, semiaquatic, and strictly aquatic) (Table 1) and are thus a sample of a broad range of ecologies and life histories. To get unbiased partitions of genetic diversity, individuals were sampled randomly from 125 and 70 different locations (table S1) in Sri Lanka and the Western Ghats of southern India, respectively. We sequenced fragments of mitochondrial DNA for each specimen and then selected one individual per unique haplotype per geographic region for further phylogenetic analysis (*15*).

Our analyses indicate that the Sri Lankan fauna is derived from an evolutionarily diverse faunal stock on the Indian mainland (*16*). However, the inferred phylogenetic trees also demonstrate that the overall limited biotic interchange has left both areas with an unexpectedly large number of endemics. For example, the Sri Lankan *Philautus* tree frogs (Fig. 2A) are the result of an extensive radiation on the island (*17*), and a small clade of deeply nested Indian tree frogs provides evidence for back

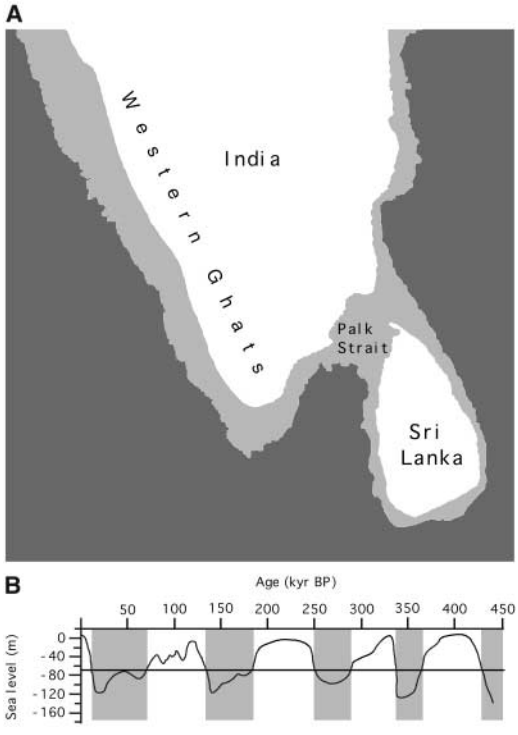


Fig. 1. (**A**) India and Sri Lanka (current outline in white) are part of the same continental shelf (light gray), which does not exceed 70 m (light gray/dark gray border) in depth. (**B**) During the past 500,000 years, sea level variations (*6*) dropping below –70 m (the horizontal line) caused Sri Lanka to be connected to India on several occasions (shaded columns) by a >100-km-broad land bridge. kyr BP, thousands of years before present.

[1]Biology Department, Unit of Ecology and Systematics, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [2]Laboratory of Evolutionary Genetics, Université Libre de Bruxelles, Code Postal, 300, Institute for Molecular Biology and Medicine, Rue Jeener and Brachet 12, B-6041 Gosselies, Belgium. [3]Department of Biology, Boston University, 5 Cummington Street, Boston, MA 02215, USA. [4]Wildlife Heritage Trust, 95 Cotta Road, Colombo 8, Sri Lanka. [5]Department of Zoology, The Natural History Museum, London SW7 5BD, UK. [6]Department of Biological Sciences, National University of Singapore, Kent Ridge, Singapore 119260, Republic of Singapore. [7]Department of Zoology, University of Kerala, Kariavattom 695581, Thiruvananthapuram, Kerala, India.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: fbossuyt@vub.ac.be

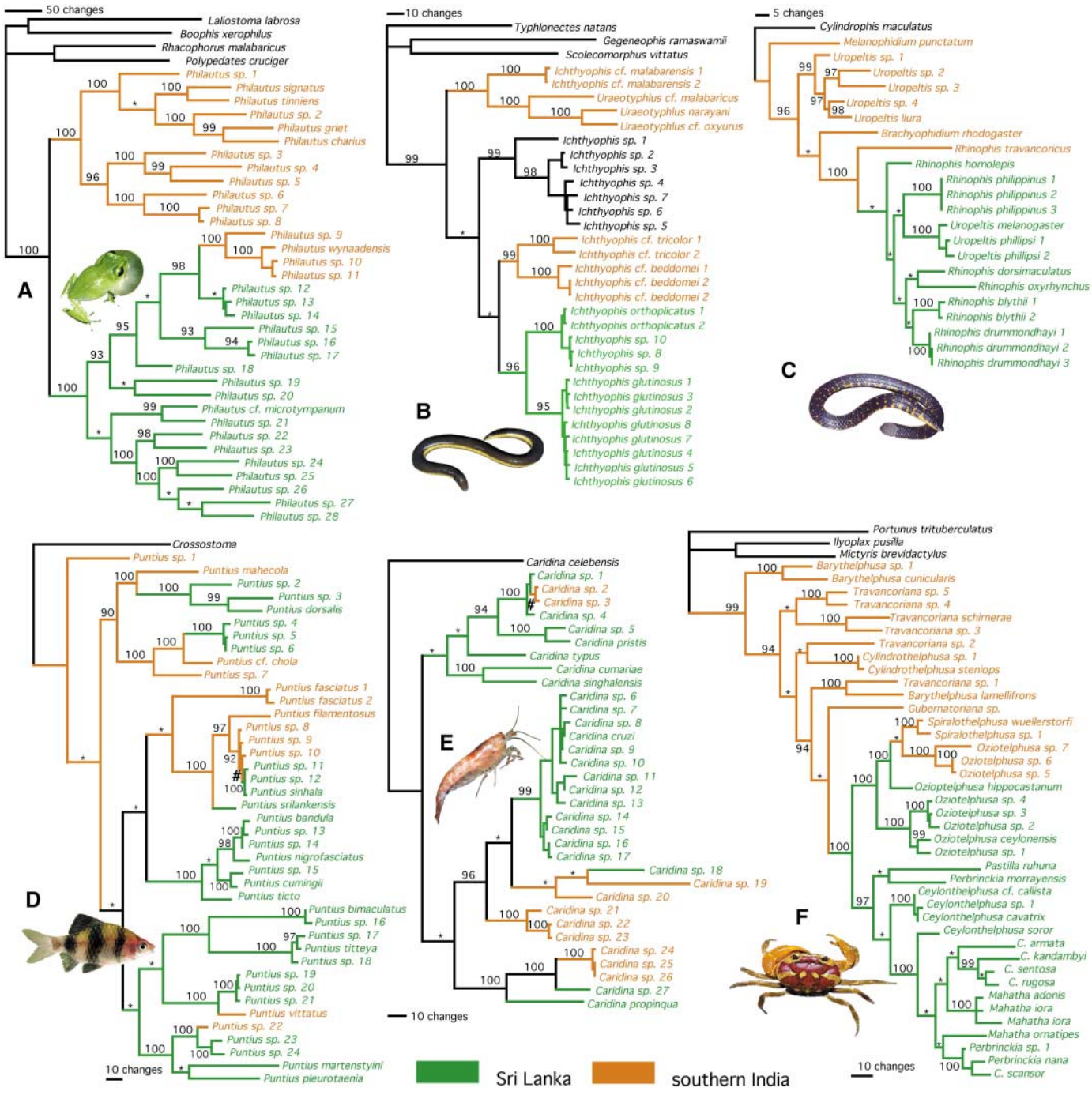


Fig. 2. Phylogenetic relationships among Indian (orange) and Sri Lankan (green) species as revealed by one of the most parsimonious trees for (**A**) tree frogs, (**B**) caecilians, (**C**) uropeltid snakes, (**D**) freshwater fishes, (**E**) freshwater shrimps, and (**F**) freshwater crabs. The strict consensus of equally parsimonious trees for each of these is shown in fig. S1. Black names represent outgroup species, except for *Ichthyophis*, which represents Southeast Asian taxa. Numbers on branches and asterisks indicate metapopulation Genetic Algorithm metaGa branch support values of ≥90% and <90%, respectively. Parsimony bootstrap values and Bayesian posterior probabilities are given in figs. S1 and S2, respectively. Numerical designations of operational taxonomic units indicate different haplotypes for mitochondrial DNA, not necessarily different species. Splits indicated with # represent recent exchanges between the mainland and the island.

dispersal of a single lineage to southern India. Similarly, our freshwater crab phylogeny revealed a radiation into several endemic genera of parathelphusids on Sri Lanka, followed by limited dispersal to India in the lowland-associated clade (*Oziotelphusa* and *Spiralothelphusa*) (Fig. 2F). In accord with morphological studies (*18*, *19*), no gecarcinucids sensu stricto were found on Sri Lanka, leaving no evidence for successful colonization of the island. The uniqueness of both sides of the Palk Strait is most noticeably illustrated by caecilians and shieldtail snakes: In both cases, all sampled island species represent endemic monophyletic groups (Fig. 2, B and C). Finally, although the pattern of limited biotic exchange is less apparent in strictly aquatic groups (Table 1), part of Sri Lanka's fish and shrimp species nevertheless form distinct clades (Fig. 2, D and E). These observations jointly indicate

Table 1. Taxa included in this study.

Taxon	Total number of specimens	Unique haplotypes	Habitat
Tree frogs	44	34	Terrestrial (including arboreal)
Caecilians	35	28	Subterranean
Uropeltid snakes	33	22	Subterranean
Freshwater fishes	51	41	Strictly aquatic
Freshwater crabs	77	40	Semiaquatic
Freshwater shrimps	44	33	Strictly aquatic

that exchange between the mainland southern Indian and insular Sri Lankan faunas has been severely restricted, despite the recurrent existence of a broad (>100-km) land bridge (*5*) during several episodes of sea level lowstands (Fig. 1B).

We used the sequence data to estimate the age of biotic exchange events (fig. S2, purple numbers) in each of the six groups. Our calculations (table S4) preclude a late Pleistocene origin for all but two splits and indicate that the corresponding events occurred before the multiple sea level lowstands of the past 500,000 years. These results are reinforced by the fact that our field surveys and phylogenetic analyses did not reveal conspecific populations in India and Sri Lanka in the four terrestrial, subterranean, and semiaquatic groups (Table 1). This was unexpected because, throughout their taxonomic history, there have been many instances in which populations on both sides of the oceanic barrier have been regarded as conspecific (*8–10*, *12*).

Our analyses show that numerous rainforest species form endemic clades, clearly identifying the Western Ghats and Sri Lanka's wet zone as distinct units. There are two possible reasons why biologists may have overlooked the differentiation between Indian and Sri Lankan faunas. First, incorrect systematic affiliations of specimens is understandable a posteriori, because our phylogenies identify homoplasy in coloration and general morphology in all groups. Second, the Sri Lankan fauna comprises a widely distributed, dry lowcountry element and a more diverse but restricted rainforest component (*20*). Because the former contains several species common to the dry zones of northern Sri Lanka and southern India that are likely Pleistocene dispersers, it has been assumed that this pattern could be generalized across the whole region.

Exact causes for the restricted dispersal between India and Sri Lanka remain speculative, but our findings highlight the importance of less conspicuous factors as important barriers to terrestrial dispersal. The faunal insularity between the wet zone of Sri Lanka and the moist forests of the Western Ghats likely results from the inability of rainforest organisms to disperse across the intervening dry lowlands. Although the climatic history of South Asia remains poorly understood, our results and the current climatic correlation between the plains of southern India and northern Sri Lanka (*21*) are possibly indicative of similar conditions during the late Pleistocene, contrary to the idea that rainforest spread onto the land bridge during periods of low sea level (*22*). Hence, montane areas and their associated climate and vegetation, rather than the present-day coastal outline, may constitute isolated islands in which the rainforest-adapted fauna has been trapped for long periods (*23*, *24*). We therefore expect that similar patterns of restricted dispersal exist elsewhere on the subcontinent, such as between opposite sides of the Palghat gap, a broad valley that traverses the southern Western Ghats. The high degree of endemicity in some species of the subcontinent is compatible with this prospect; tree frogs, uropeltids, and freshwater crabs, for example, include point endemics with distributions of often just a few square kilometers (*25–27*). Thus, treating the Western Ghats and Sri Lanka as a single hotspot carries with it the danger of overlooking strong biogeographic structure within this region (*28*, *29*). Conservation management of the Indian subcontinent will benefit from further characterization of the heterogeneity of biodiversity down to more local scales.

References and Notes

1. G. G. Gillespie, G. K. Roderick, *Annu. Rev. Entomol.* **47**, 595 (2002).
2. R. H. MacArthur, E. O. Wilson, *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, NJ, 1967).
3. P. J. Darlington, *Zoogeography: The Geographical Distribution of Animals* (Wiley, New York, 1957).
4. C. D. Schubart, R. Diesel, S. B. Hedges, *Nature* **393**, 363 (1998).
5. T. Somasekaram, Ed., *Atlas of Sri Lanka* (Arjuna Consulting, Dehiwela, Sri Lanka, 1997).
6. E. J. Rohling *et al.*, *Nature* **394**, 162 (1998).
7. G. G. Vaz, *Curr. Sci.* **79**, 228 (2000).
8. P. Kirtisinghe, *The Amphibian Fauna of Ceylon* (self-published, Colombo, Sri Lanka, 1957).
9. R. F. Inger, H. B. Shaffer, M. Koshy, R. Bakde, *J. Bombay Nat. Hist. Soc.* **81**, 551 (1984).
10. M. A. Smith, *Serpentes* (*Fauna of British India, Reptilia and Amphibia*, Taylor & Francis, London, 1943), vol. 3.
11. R. Pethiyagoda, *Freshwater Fishes of Sri Lanka* [Wildlife Heritage Trust (WHT), Colombo, Sri Lanka, 1991].
12. R. Bott, *Abh. Senckenb. Naturforsch. Ges.* **526**, 1 (1970).
13. A. R. Wallace, *The Geographical Distribution of Animals* (Macmillan, London, 1876).
14. N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, J. Kent, *Nature* **403**, 853 (2000).
15. Materials and methods are available as supporting material on *Science* Online.
16. The geographic origin and/or direction of dispersal of a clade can only be established if sufficient sampling is available from the whole distribution area. As such, a single mainland origin of Sri Lankan lineages is currently indicated in three of the six examined groups because of their nested position with respect to Indian and/or Asian lineages: caecilians and uropeltid snakes (both indicated by our analyses) and *Philautus* tree frogs [not evident from our tree, but shown in (*17*)]. A mainland origin for Sri Lankan clades is not contradicted in the three other groups, but will only be unambiguously confirmed when more inclusive phylogenies are available for these groups.
17. M. Meegaskumbura *et al.*, *Science* **298**, 379 (2002).
18. P. K. L. Ng, F. W. M. Tay, *Zeylanica* **6**, 113 (2001).
19. R. Bott, *Ark. Zool.* **22**, 627 (1970).
20. F. R. Senanayayake, M. Soulé, J. W. Senner, *Nature* **265**, 351 (1977).
21. G. B. Pan, K. Rupa Kumar, *Climates of South Asia* (Wiley, New York, 1997).
22. W. Erdelen, C. Preu, in *Vegetation and Erosion*, J. B. Thornes, Ed. (Wiley, Chichester, UK, 1990), pp. 491–504.
23. J. E. Cadle, H. C. Dessauer, C. Gans, D. F. Gartside, *Biol. J. Linn. Soc.* **40**, 293 (1990).
24. C. Moritz, L. Joseph, M. Cunningham, C. J. Schneider, in *Tropical Rainforest Remnants: Ecology, Management, and Conservation of Fragmented Communities*, W. F. Laurance, R. O. Bierregaard, Eds. (Univ. of Chicago Press, Chicago, 1997), pp. 442–465.
25. R. J. R. Daniels, *Curr. Sci.* **81**, 240 (2001).
26. R. Pethiyagoda, K. Manamendra-Arachchi, *Occas. Pap. Wildl. Heritage Trust* **2**, 1 (1998).
27. P. K. L. Ng, *J. S. Asian Nat. Hist.* **1**, 129 (1995).
28. J. R. Prendergast, R. M. Quinn, J. H. Lawton, B. C. Eversham, D. W. Gibbons, *Nature* **365**, 335 (1993).
29. A. S. L. Rodrigues *et al.*, *Nature* **428**, 640 (2004).

30. We thank the Forest Department and the Department of Wildlife Conservation, Sri Lanka, for research permission; J. Spinks, S. Loader, and S. Meegaskumbura for lab work; the Louisiana State University Museum of Natural Science's Collection of Genetic Resources for tissues; D. Raheem, Y. Mapatuna, F. Naggs (U.K. Darwin Initiative grant no. 162/08/214), S. Kankanam-Gamage, K. Wewelwala, S. Batuwita, and R. Wickramatilleke for fieldwork; and A. Captain, S. Thakur, and C. Luckhaup for photographs. Sequences have been deposited at GenBank under accession nos. AY700937 to AY700990 (caecilians); AY700999 to AY701021 and AY701030 to AY701052 (snakes); AY706108 to AY706131 and AY708128 to AY708196 (frogs); AY708197 to AY708278 (fishes); AY708052 to AY708091 (crabs); and AY708092 to AY708127 (shrimps). F.B. is a postdoctoral researcher and K.R. an aspirant at the Fonds voor Wetenschappelijk Onderzoek (FWO)–Vlaanderen. Supported by FWO–Vlaanderen grant nos. G.0056.03 and 1.5.039.03 (F.B.), Vrije Universiteit Brussel–Onderzoekrsaad (F.B. and K.R.); Fonds National de la Recherche Scientifique, the "Communauté Française de Belgique" (Action de Recherches Concertées no. 11649/20022770); the Walloon Region (BioRobot-Initiative no. 114840) (M.C.M.); Boston Univ. and NSF grant no. DEB9977072 (C.J.S. and M.M.); Leverhulme Trust grant no. F/00696/F (D.J.G and M.W.); and WHT Sri Lanka (R.P., M.M., M.M.B., and K.M-A).

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/479/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 to S4
References and Notes

11 May 2004; accepted 1 September 2004

Regulation of Gene Expression by a Metabolic Enzyme

David A. Hall,[1] Heng Zhu,[2*] Xiaowei Zhu,[3] Thomas Royce,[1] Mark Gerstein,[1] Michael Snyder[1,2†]

Gene expression in eukaryotes is normally believed to be controlled by transcriptional regulators that activate genes encoding structural proteins and enzymes. To identify previously unrecognized DNA binding activities, a yeast proteome microarray was screened with DNA probes; Arg5,6, a well-characterized mitochondrial enzyme involved in arginine biosynthesis, was identified. Chromatin immunoprecipitation experiments revealed that Arg5,6 is associated with specific nuclear and mitochondrial loci in vivo, and Arg5,6 binds to specific fragments in vitro. Deletion of Arg5,6 causes altered transcript levels of both nuclear and mitochondrial target genes. These results indicate that metabolic enzymes can directly regulate eukaryotic gene expression.

Although gene products with enzymatic and structural functions have been known to regulate gene expression through indirect mechanisms (*1*, *2*), whether enzymatic proteins can directly control gene expression has not been extensively investigated. Butow and co-workers found that a protein involved in isoleucine synthesis, Ilv5, is associated with mitochondrial DNA in vivo (*3*, *4*). However, it is not known whether this enzyme was directly associated with DNA, bound specific sequences, or had a direct role in affecting the expression of specific genes. Zheng *et al*. (*5*) reported that glyceraldehyde-3-phosphate dehydrogenase (GAPDH) serves as a coactivator to regulate the expression of histone H2B. GAPDH does not directly associate with DNA and may have an indirect role; moreover, its ability to regulate other loci and thus serve as a global regulator was not investigated.

To further understand mechanisms of gene regulation, we used proteome arrays and chromatin immunoprecipitation (ChIP/chip) to identify previously unrecognized DNA binding activities (Fig. 1) (*6*). Protein microarrays containing nearly all of the proteins of yeast were spotted in duplicate onto microscope slides (*7*) and probed with either single-stranded or double-stranded yeast genomic DNA labeled with Cy3; 6 and 10 arrays were probed with single-stranded DNA and double-stranded DNA probes, respectively. Of the >200 total identified DNA binding proteins that reproducibly bound double- or single-stranded DNA, 84 proteins bound single-stranded DNA, 58 bound double-stranded DNA, and 131 bound both (see table S1 for a detailed list).

Many (~50%) of the proteins identified are expected to bind DNA based on their known functions (e.g., transcription factors). However, a large number were either not known or would not be suspected to have DNA binding activity based on available information. These latter proteins could (i) represent previously unrecognized DNA binding proteins, (ii) be associated with DNA binding proteins that copurified with the protein present on the proteome array, or (iii) bind DNA nonspecifically in vitro but not in vivo, and thus be artifacts.

We tested whether eight proteins (Arg5,6, Dig2, Mtw1, Yrb2, Akl1, Yer152c, Lrg1, and Rub1), which have not previously been reported to bind DNA, were specifically associated with chromosomal DNA in vivo by immunoprecipitation of the protein and probing of genomic DNA microarrays [ChIP/chip (*8*)]. Each candidate DNA binding protein was first tagged at its C terminus with 13 copies of c-Myc epitopes (13Xmyc) by insertion of the epitope coding sequences at the endogenous locus. Immunoblot analysis revealed that each protein migrated at its expected size. The exception was Arg5,6, which is normally cleaved into two peptides, an N-terminal N-acetyl-gamma-glutamyl phosphate reductase (Arg6) and C-terminal acetylglutamate kinase (Arg5); the tagged protein migrated at ~70 kD, the size expected of Arg5::13Xmyc (Fig. 2A). The cells were treated with formaldehyde to cross-link protein and DNA, and the cells lysed and sonicated. The epitope-tagged protein was immunoprecipitated and the associated DNA purified and labeled with Cy5. As a control, DNA from an untagged strain was prepared in parallel and labeled with Cy3. The two probes were mixed and used to probe a yeast genomic DNA array containing the intergenic regions and the entire mitochondrial genome. To ensure lack of bias, experiments were performed blindly with respect to the particular protein analyzed. Immunoprecipitates of five epitope-tagged yeast proteins (Yrb2, Akl1, Yer152c, Lrg1, and Rub1) did not show enrichment of any chromosomal loci over the untagged strains. These proteins either do not associate with DNA in cells grown in rich medium or they associate with DNA nonspecifically; they were not pursued further.

Three proteins—Mtw1, Dig2, and Arg5,6—were found to be associated with specific DNA regions in vivo. Mtw1, a kinetochore protein immunoprecipitated nearly all of the yeast centromeres (*9*), and Dig2, a negative regulator of the Ste12 transcription factor (*10*) bound a number of loci. These proteins either bind DNA directly or are associated with a DNA binding component and will be pursued as part of a separate study. In four independent ChIP/chip experiments, Arg5 was found to be reproducibly associated with a number of DNA fragments [with the use of a high stringency cut-off ($P < 0.001$) (Fig. 2B) (*6*)]. Arg5,6 encodes two mitochon-

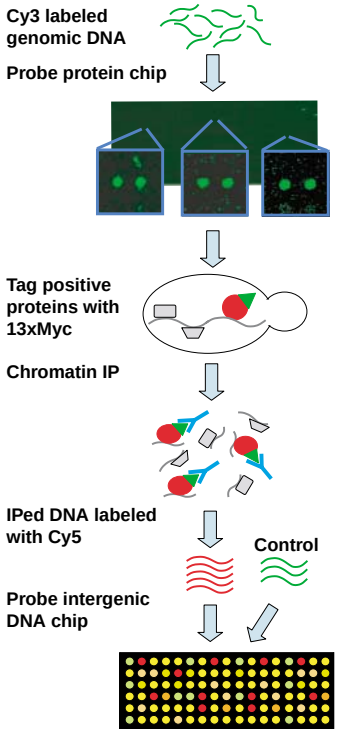


Fig. 1. A microarray containing 6500 protein preparations of 5800 different yeast proteins spotted in duplicate (*7*) was probed with genomic yeast double-stranded DNA labeled with Cy3. Positives were identified, and eight were tested for association with specific chromosomal and mitochondrial regions with the use of ChIP/chip.

[1]Department of Molecular Biophysics and Biochemistry, [2]Department of Molecular, Cellular, and Developmental Biology, [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520–8005, USA.

*Present address: Department of Pharmacology and the High Throughput Biology Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

†To whom the correspondence should be addressed. E-mail: michael.snyder@yale.edu

drial enzymes, Arg6 and Arg5, which mediate two key steps in the biosynthesis of ornithine, which is a precursor to arginine. Neither of the Arg5 or Arg6 gene products was known previously to regulate gene expression, but nitrogen metabolism is known to be closely linked with mitochondrial function and ornithine synthesis (*11*, *12*).

Arg5 bound 22 fragments which contained primarily mitochondrial loci, including the 15*S* ribosomal DNA (rDNA), *COX3*, three regions of *COB1*, and 10 regions within *COX1* (*13*–*15*); the latter two genes contain introns (table S2 and Fig. 2C). Arg5 is also associated with several nuclear loci including regions adjacent to *PUF4*, *PHO23*, and *THI13*, indicating that this protein likely resides in both cellular compartments.

We next performed ChIP experiments with tagged and untagged strains and tested for enrichment of specific loci in the tagged strain by standard polymerase chain reaction (PCR) assays (*8*). Fragments were specifically enriched for many nuclear and mitochondrial loci immunoprecipitated from the tagged strain relative to those from the untagged strain; these include four *COX1* regions, *COB1*, *COX3*, *YOR352w*, *PHO23*, *THI13*, and *PUF4* (Fig. 2C). Control fragments (*SUN4*, *CDC7*, and *CHS5*) were not enriched in the Arg5-tagged immunoprecipitates. Thus, Arg5 is associated with mitochondrial and nuclear DNA in vivo.

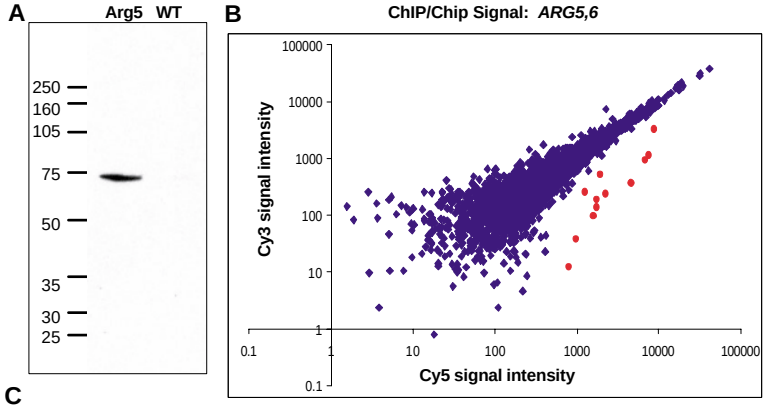
To determine whether Arg5,6 can bind DNA directly, Arg5,6 was overproduced in yeast as a glutathione *S*-transferase (GST) fusion, purified with the use of stringent conditions (in the presence of 0.5 M NaCl) and tested for binding to two *COX1* targets, a fragment at the 5′ end *COX1* and another in the first exon, with the use of in vitro "gel-shift" assays. Increasing amounts of Arg5,6 were mixed with the *COX1* targets or negative control DNAs [Epstein-Barr virus nuclear antigen (EBNA) or three mitochondrial fragments that do not bind Arg5,6] and separated in a native polyacrylamide gel. As an additional control, GST alone was added to the *COX1* DNA. As shown in Fig. 3A, *COX1* DNA forms a slower migrating complex in the presence of Arg5,6, but not in the presence of GST. The negative control DNAs (EBNA and other mitochondria segments) do not form a complex (Fig. 3C) (*16*). Thus, Arg5,6 can specifically associate with *COX1* DNA in vitro.

To determine the site bound by Arg5,6, the sequences of the in vivo targets were compared and a strong common motif was identified ($P < 10^{-5}$; Fig. 3B) (*17*). Incubation of the Arg5,6 protein with a labeled double-stranded oligonucleotide containing this sequence from the *COX1* region resulted in the formation of a complex. The complex was not formed with a randomized version of this sequence or in the presence of unlabeled competitor oligonucleotide. Thus, Arg5,6 associates with a specific DNA sequence. This sequence is 78% GC, unlike mitochondrial DNA, which is 83% AT. Whether Arg5 or Arg6 binds this sequence in vitro is not clear; however, because Arg5 associates with specific loci in vivo, it is the most probable candidate for this activity.

Mutant analysis revealed no morphological or growth defects in *arg5,6*Δ cells grown in rich medium, nitrogen-limiting medium, and medium containing glycerol as the primary carbon source. The distribution and intensities of mitochondrial DNA staining in cells incubated in the same conditions were also identical in wild-type and *arg5,6*Δ cells.

To determine whether Arg5,6 has a role in regulating gene expression, we used real-time PCR (RT-PCR) to quantify the level of *COX1*, *COB1*, and *COX3* mRNAs and the mRNA levels of nuclear targets *PUF4*, *YOR352w*, and *YHL045w* in wild-type and *arg5,6*Δ strains. Cells were examined in rich medium and in medium limited for nitrogen or amino acids; in the absence of amino acids, the levels of many mitochondrial enzymes are induced (*18*). *arg5,6*Δ strains grown in medium lacking amino acids exhibited significantly lower levels (3.5- to 6-fold) of the *COX1* mitochondrial and the *YOR352W* nuclear transcripts as compared with those of wild-type cells; 2.2-fold lower *PUF4* mRNA levels were observed in the same conditions (Fig. 4). Similar but weaker effects were observed when cells were grown in medium limited for nitrogen. Only a modest effect was observed for different messages when cells are grown in rich medium. The mRNA levels of three genes—*ACT1*, *COX2*, and 21*S* ribosomal RNA (the latter two are mitochondrial)—that are not targets of Arg5,6 are not affected. Thus, Arg5,6 is important for controlling the levels of specific mitochondrial and nuclear transcripts in cells lacking amino acids or limited for nitrogen; whether Arg5 or Arg6 is responsible for this activity is not resolved by these experiments.

The presence of Arg5,6 binding sites in the intronic regions of *COX1* and *COB1* raises the possibility that Arg5,6 might also play a role in RNA processing. We therefore designed real-time RT-PCR probes to monitor the levels of unprocessed (i.e., intronic) *COX1* and *COB1* messages in *arg5,6*Δ cells grown in the media described above. As shown in Fig. 4, unprocessed transcript levels for *COX1* are slightly affected by the presence of Arg5,6, whereas those for *COB1* are significantly affected. Thus, these data indicate that Arg5,6 affects the levels of unprocessed RNA for *COB1*, as expected for a transcriptional regulator; processed *COX1* levels are more significantly affected than unprocessed messages, raising the possibility that Arg5,6 might also directly or indirectly participate in *COX1* RNA processing. Because Arg5,6 regulates the *COB1* locus, which encodes a maturase in its precursor RNA that processes *COX1*



C

Fragment	15S rRNA2	PC1 COB	PC2 COX1	PC4 COX1	PC6 COX1	PC8 COX1	PC2 YOR352W	PC2 PHO23	PC1 THI13	PC YHL045W	PC PUF4	Negative Control
Fold Enrichment	10.1±8.6	3.9±2.0	22.0±7.9	3.7±2.2	13.8±12.0	4.5±0.2	8.5±3.6	8.4±4.1	>30	2.0±0.4	3.8±2.3	1.0±0.1

Fig. 2. (**A**) Immunoblot analysis of Arg5. Proteins were prepared from untagged wild-type cells and an Arg5,6 13Xmyc strain containing 13 copies of Myc coding sequences at the C terminus. Immunoblots were prepared and probed with antibodies to Myc. Left lane (+): the Arg5::13Xmyc strain. Right lane (–): the untagged strain. WT, wild type. Migration of molecular weight markers in kD is indicated on the left. (**B**) Identification of loci bound by Arg5. Arg5::13Xmyc-bound DNA was purified by immunoprecipitation and labeled with Cy5; control samples purified from an untagged strain were labeled with Cy3. The DNA samples were mixed and used to probe a DNA microarray containing the intergenic regions of yeast. Fragments exhibiting a strong enrichment ($P < 0.001$) in the tagged strain relative to that of the untagged strains in four separate experiments were identified with the use of ExpressYourself (*8*). The plot of Cy5/Cy3 ratios for each fragment is present; significantly enriched fragments are indicated in red. (**C**) Table summarizing the PCR confirmation of 11 regions that bind Arg5. Fold-enrichments of three separate experiments are presented. Enrichment was observed for *COX1*, *COB1*, *YOR352w*, *PHO23*, *THI3*, *PUF4*, and *YH045w* loci but not control loci. Fold enrichment data are shown as mean ± standard error.

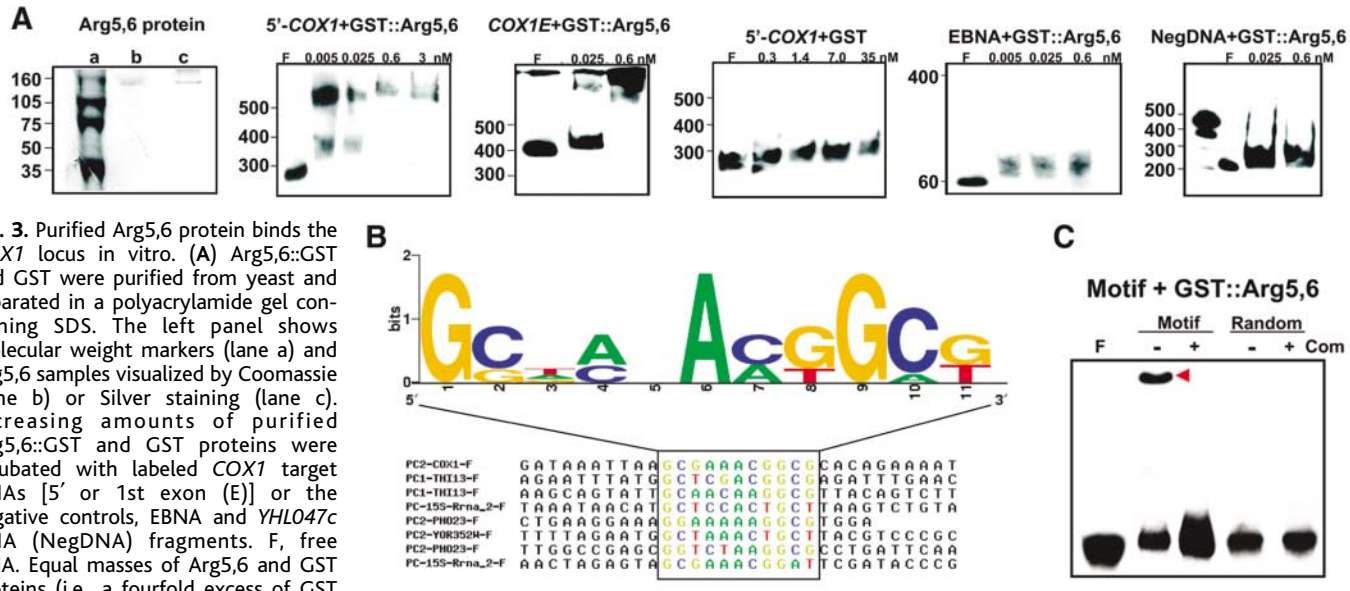


Fig. 3. Purified Arg5,6 protein binds the *COX1* locus in vitro. (**A**) Arg5,6::GST and GST were purified from yeast and separated in a polyacrylamide gel containing SDS. The left panel shows molecular weight markers (lane a) and Arg5,6 samples visualized by Coomassie (lane b) or Silver staining (lane c). Increasing amounts of purified Arg5,6::GST and GST proteins were incubated with labeled *COX1* target DNAs [5′ or 1st exon (E)] or the negative controls, EBNA and *YHL047c* DNA (NegDNA) fragments. F, free DNA. Equal masses of Arg5,6 and GST proteins (i.e., a fourfold excess of GST molecules) were tested at the indicated concentrations. Arg5,6 fusion proteins specifically bound to *COX1* DNAs, whereas GST did not. Arg5,6 did not bind to the EBNA DNA or other control DNA fragments (*16*). (**B**) Consensus motif derived from targets confirmed by PCR. (**C**) Complex formation of Arg5,6 in the presence of the labeled *COX1* double-stranded oligonucleotide containing three copies of the motif (lane 2), but not in the presence of 200-fold excess of unlabeled competitor double-stranded oligonucleotide or in the presence of a randomized *COX1* motif double-stranded oligonucleotide. Com, unlabeled competitor DNA.

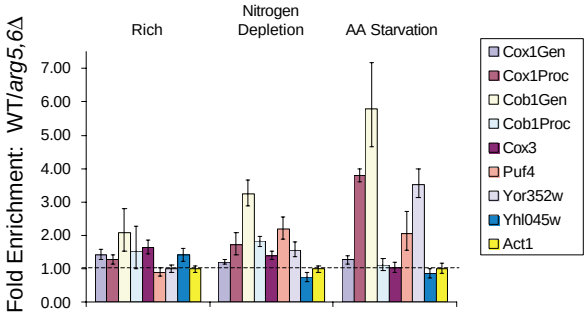


Fig. 4. Reduced levels of specific mitochondrial and nuclear transcripts in cells lacking Arg5,6. Wild-type and *arg5,6*Δ cells were grown in either rich medium or media limited for nitrogen or lacking amino acids (AA). The transcripts levels of several Arg5,6 targets were quantified with the use of real-time RT-PCR. The transcript ratios were normalized to actin and relative transcript levels in wild-type relative to *arg5,6*Δ cells are presented.

RNA (*19*), the reduced level of *COB1* is expected to decrease *COX1* RNA processing.

The association of Arg5 within the *COX1*, *COB1*, *COX3*, and 15*S* rDNA genes, rather than just at the initiation sites, raises the possibility that Arg5,6 plays a role in either transcription elongation or RNA processing. However, it is likely that Arg5,6 also affects transcription because many of its target loci lack introns (e.g., *COX3*, *PUF4*, *PHO23*, and *THI13*). Thus, Arg5,6 may have a role in the regulation of both transcriptional and post-transcriptional processes. Coupling of transcription and RNA processing has been reported in eukaryotes previously (*20*, *21*).

Although the presence of Arg5 in the mitochondria has been well described (*22*), this protein has not been previously found in the nucleus by ChIP. We presume that the nuclear levels of Arg5 are low because we could not detect Arg5::13Xmyc by indirect immunofluorescence.

Nitrogen metabolism has been linked to mitochondrial function in yeast (*11*, *12*). Discovery of a protein involved in ornithine biosynthesis, which is linked to nitrogen metabolism, which regulates the activity of mitochondrial gene expression, provides a direct link for how this might occur.

Several examples of multifunctional proteins have been reported previously. Butow and co-workers identified proteins associated with mitochondrial DNA and found tricarboxylic acid cycle proteins and several involved in amino acid biosynthesis (*23*). These proteins were demonstrated to be important for mitochondrial stability. It is plausible that many proteins have multiple roles in vivo and unbiased proteomic approaches such as those used here (proteome chips and chIP/chip) will be required to fully identify the plethora of functions of eukaryotic proteins.

References and Notes

1. B. M. Turner, *Nature Cell Biol.* **5**, 390 (2003).
2. D. L. Black, *Annu. Rev. Biochem.* **72**, 291 (2003).
3. J. M. Bateman, P. S. Perlman, R. A. Butow, *Genetics* **161**, 1043 (2002).
4. J. M. Bateman, M. Iacovino, P. S. Perlman, R. A. Butow, *J. Biol. Chem.* **277**, 47946 (2002).
5. L. Zheng, R. G. Roeder, Y. Luo, *Cell* **114**, 255 (2003).
6. Materials and methods are available as supporting material on *Science* Online.
7. H. Zhu *et al.*, *Science* **293**, 2101 (2001).
8. C. E. Horak *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2924 (2002).
9. B. A. Pinsky, S. Y. Tatsutani, K. A. Collins, S. Biggins, *Dev. Cell* **5**, 735 (2003).
10. A. B. Kusari, D. M. Molina, W. Sabbagh, Jr., C. S. Lau, L. Bardwell, *J. Cell Biol.* **164**, 267 (2004).
11. A. Abadjieva, K. Pauwels, P. Hilven, M. Crabeel, *J. Biol. Chem.* **276**, 42869 (2001).
12. J. M. Guillamon, N. A. van Riel, M. L. Giuseppin, C. T. Verrips, *FEMS Yeast Res.* **1**, 169 (2001).
13. S. Naithani, S. A. Saracco, C. A. Butler, T. D. Fox, *Mol. Biol. Cell* **14**, 324 (2003).
14. F. H. MacIver, I. W. Dawes, C. M. Grant, *Curr. Genet.* **31**, 119 (1997).
15. B. J. Hicke, E. L. Christian, M. Yarus, *EMBO J.* **8**, 3843 (1989).
16. D. A. Hall *et al.*, data not shown.
17. T. L. Bailey, M. Gribskov, *J. Comput. Biol.* **5**, 211 (1998).
18. K. Natarajan *et al.*, *Mol. Cell. Biol.* **21**, 4347 (2001).
19. W. M. Schmidt, R. J. Schweyen, K. Wolf, M. W. Mueller, *J. Mol. Biol.* **243**, 157 (1994).
20. Y. Hirose, J. L. Manley, *Genes Dev.* **14**, 1415 (2000).
21. S. H. Ahn, M. Kim, S. Buratowski, *Mol. Cell* **13**, 67 (2004).
22. A. Sickmann *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13207 (2003).
23. B. A. Kaufman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7772 (2000).

24. We thank G. Jona, D. Gelperin, M. Smith, and L. Kung for manuscript comments and J. Ptacek and L. Meng for experimental assistance. D.H. was supported by an NIH training grant and H.Z. by a fellowship from the Damon Runyon–Walter Winchell Cancer Research Foundation. This research was supported by grants from the NIH.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/482/DC1

Materials and Methods

Figs. S1 and S2

Tables S1 and S2

13 February 2004; accepted 5 August 2004

Prevention of Vaginal SHIV Transmission in Rhesus Macaques Through Inhibition of CCR5

Michael M. Lederman,[1]*† Ronald S. Veazey,[2]* Robin Offord,[3]* Donald E. Mosier,[4] Jason Dufour,[2] Megan Mefford,[2] Michael Piatak Jr.,[5] Jeffrey D. Lifson,[5] Janelle R. Salkowitz,[1] Benigno Rodriguez,[1] Andrew Blauvelt,[6]‡ Oliver Hartley[3]

Topical agents, such as microbicides, that can protect against human immunodeficiency virus (HIV) transmission are urgently needed. Using a chimeric simian/human immunodeficiency virus (SHIV SF162), which is tropic for the chemokine receptor CCR5, we report that topical application of high doses of PSC-RANTES, an amino terminus–modified analog of the chemokine RANTES, provided potent protection against vaginal challenge in rhesus macaques. These experimental findings have potentially important implications for understanding vaginal transmission of HIV and the design of strategies for its prevention.

Because the vast majority of HIV infections are acquired via transmission across mucosal surfaces, strategies to prevent mucosal transmission are urgently needed. Unfortunately, the mechanisms whereby HIV gains entry at mucosal sites, especially vaginal sites of infection, are incompletely understood. Thus, there is no uniform agreement regarding the critical host cellular and molecular targets during infection after vaginal exposure, and resolution of these issues is needed for the design of plausible microbicide strategies to prevent mucosally acquired HIV infection.

The chemokine receptor CCR5 serves as an essential cofactor for HIV entry and acquisition of infection. Thus, persons whose cells lack surface CCR5 expression because of mutation are almost completely protected from acquiring HIV infection (*1*–*5*). Furthermore, viruses that utilize CCR5 predominate in early stages of mucosal transmission (*6*, *7*), which suggests that mucosal transmission may selectively involve CCR5. Hence, inhibition of CCR5 has been proposed as a possible "microbicide" strategy for prevention of HIV infection.

[1]Department of Medicine, Case Western Reserve University, University Hospitals, 2061 Cornell Road, Cleveland, OH 44106, USA. [2]Tulane National Primate Research Center, 18703 Three Rivers Road, Covington, LA 70433, USA. [3]Department of Structural Biology and Bioinformatics, Faculty of Medicine, University of Geneva, 1211 Geneva 4, Switzerland. [4]Scripps Research Institute, 10550 Torrey Pines Road, La Jolla, CA 92037, USA. [5]AIDS Vaccine Program, SAIC Frederick, Inc., National Cancer Institute, Frederick, MD 21702-1201, USA. [6]Dermatology Branch, National Cancer Institute, Bethesda, MD 20892–1908, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: MXL6@case.edu
‡Present address: Department of Dermatology, Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97239, USA.

However, HIV is able to use other host cell factors that are present at mucosal sites to achieve or to facilitate infection (*6*–*12*). These findings have led to some debate about the importance of CCR5 for infection across mucosae, as well as concern that targeting CCR5 alone may be inadequate to prevent transvaginal HIV transmission (*13*).

We previously described the synthesis of an amino terminus–modified form of the chemokine RANTES, the aminooxypentane oxime of [glyoxylyl[1]]RANTES [2-68], known as AOP-RANTES (*14*). This compound is significantly more potent at inhibiting HIV-1 replication than the parent chemokine. Subsequently, a series of amino-terminally modified RANTES analogs have been developed and tested (*15*–*20*) in an effort to improve potency and durability of HIV inhibitory activity.

AOP-RANTES blocked in vitro propagation of multiple CCR5-using HIV isolates representing clades A to F (*17*) with inhibitory concentrations in the nanomolar range. Pretreatment of hu-PBL-SCID chimeras (mice with severe combined immunodeficiency disease reconstituted with human peripheral blood lymphocytes) with another more potent RANTES analog, N^{α}-(*n*-nonanoyl)-*des*-Ser[1]-RANTES (known as NNY-RANTES) protected animals from parenteral HIV challenge, although in some, escape with both R5- and X4-using viruses was demonstrated (*15*).

A third analog, N^{α}-(*n*-nonanoyl)-*des*-Ser[1]-[L-thioproline[2], L-α-cyclohexyl-glycine[3]]RANTES (PSC-RANTES) represents a new RANTES analog chemically identical to native RANTES except for the substitution of a nonanoyl group, thioproline, and cyclohexylglycine for the first three N-terminal amino acids of the native protein (Fig. 1). PSC-RANTES has more potent in vitro antiviral activity than earlier analogs, with inhibitory concentrations for some HIV-1 isolates in the picomolar range (*21*). The induction of receptor internalization and down-modulation on binding by chemokines is believed to play a major role in their anti-HIV action (*22*), although some contribution from competitive binding cannot be excluded (*23*, *24*). Our RANTES analogs show a particularly enhanced capacity to induce such internalization and down-modulation, and this may be the basis of their potent anti-HIV activity (*18*, *25*, *26*).

We first confirmed that PSC-RANTES inhibited propagation of the SHIV SF162 R5-tropic virus in rhesus peripheral blood mononuclear cells (PBMCs) and completely blocked SHIV SF162 replication, with median inhibitory concentration (IC_{50}) values in the subnanomolar range [Fig. 2; (*27*)]. Furthermore, PSC-RANTES caused down-modulation of macaque CCR5 (Fig. 2). After only 15 min of exposure, the decrease in CCR5 levels on both $CD4^-$ and $CD4^+$ peripheral blood cells was already maximal [>90%; (*28*)].

To examine PSC-RANTES' ability to prevent acquisition of SHIV infection at a mucosal site, 30 progesterone-treated (*27*) adult female rhesus macaques were pretreated with 4 ml PSC-RANTES at the indicated concentrations or with phosphate-buffered saline (PBS). The animals were subsequently challenged with a high multiplicity [300 $TCID_{50}$ (median tissue culture infectious dose)] of SHIV SF162 and monitored for up to 24 weeks for the development of plasma viremia (*29*). All five animals treated with the highest dose (1 mM) of PSC-RANTES were protected from SHIV infection, with no detectable viremia for the entire duration of follow-up (Fig. 3). Lower doses also proved protective, with four out of five animals treated with 330 μM and three out of five treated with 100 μM PSC-RANTES also showing protection from infection. One of five animals treated with

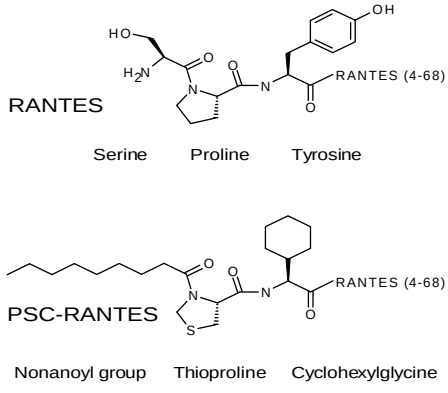


Fig. 1. Structure of native RANTES and PSC-RANTES.

33 μM PSC-RANTES and two of five animals treated with 10 μM or less were protected. Overall, 12 out of 15 animals (80%) pretreated with ≥100 μM were protected from infection, whereas only 4 out of 15 animals (27%) pretreated with <100 μM or placebo showed protection [P = 0.009, Fisher's exact test; risk ratio (RR) = 0.27, 95% confidence interval (CI), 0.09 to 0.78 (*27*)]. There was a significant dose-effect relationship when considering the whole range of dosing levels (P = 0.0048, Cochran-Armitage exact trend test). Based on exact logistic regression modeling, we estimated that in this system a 10-fold increase in PSC-RANTES dose was associated with an odds ratio (OR) of infection of 0.39 [95% CI, 0.17 to 0.82], (P = 0.035).

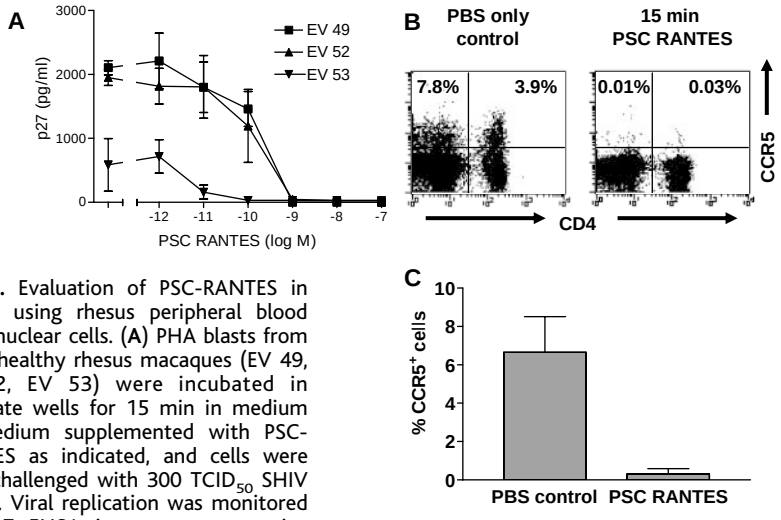


Fig. 2. Evaluation of PSC-RANTES in assays using rhesus peripheral blood mononuclear cells. (**A**) PHA blasts from three healthy rhesus macaques (EV 49, EV 52, EV 53) were incubated in triplicate wells for 15 min in medium or medium supplemented with PSC-RANTES as indicated, and cells were then challenged with 300 $TCID_{50}$ SHIV SF162. Viral replication was monitored by p27 ELISA in supernatant twice weekly and found to peak at day 11. Data shown are mean p27 levels at day 11. Error bars indicate SEM. (**B**) Whole blood was incubated with 10 μM PSC-RANTES in PBS or PBS control for 15 min. Surface expression of CCR5 was examined by flow cytometry on $CD4^-$ and $CD4^+$ peripheral blood cells. (**C**) Summary mean (standard deviation) percentages of peripheral blood cells expressing CCR5 in five separate experiments in the presence of PBS control or PSC-RANTES.

Plasma samples obtained at week 11 from all protected animals and four infected animals were tested for the presence of antibodies to simian immunodeficiency virus (SIV) proteins by Western blot. All four infected animal plasmas had strong bands corresponding to p17, p27, p55, and p66, whereas no positive bands were found in any of the protected animals (*27*). Plasma levels of PSC-RANTES were measured in samples obtained at intervals (1 hour, 4 hours, and 24 hours) after intravaginal administration of the maximum concentration used (1 mM). PSC-RANTES was undetectable (lower than the 300 pM limit of assay sensitivity) in all samples tested (*30*).

We have shown that PSC-RANTES, targeting CCR5 alone, protected rhesus macaques from intravaginal exposure to a chimeric SHIV containing an R5-tropic envelope of HIV-1. Thus, pursuing a strategy that targets this receptor seems reasonable for development. However, the concentrations used in the highest dose group (1 mM), exceed by orders of magnitude the subnanomolar IC_{50}

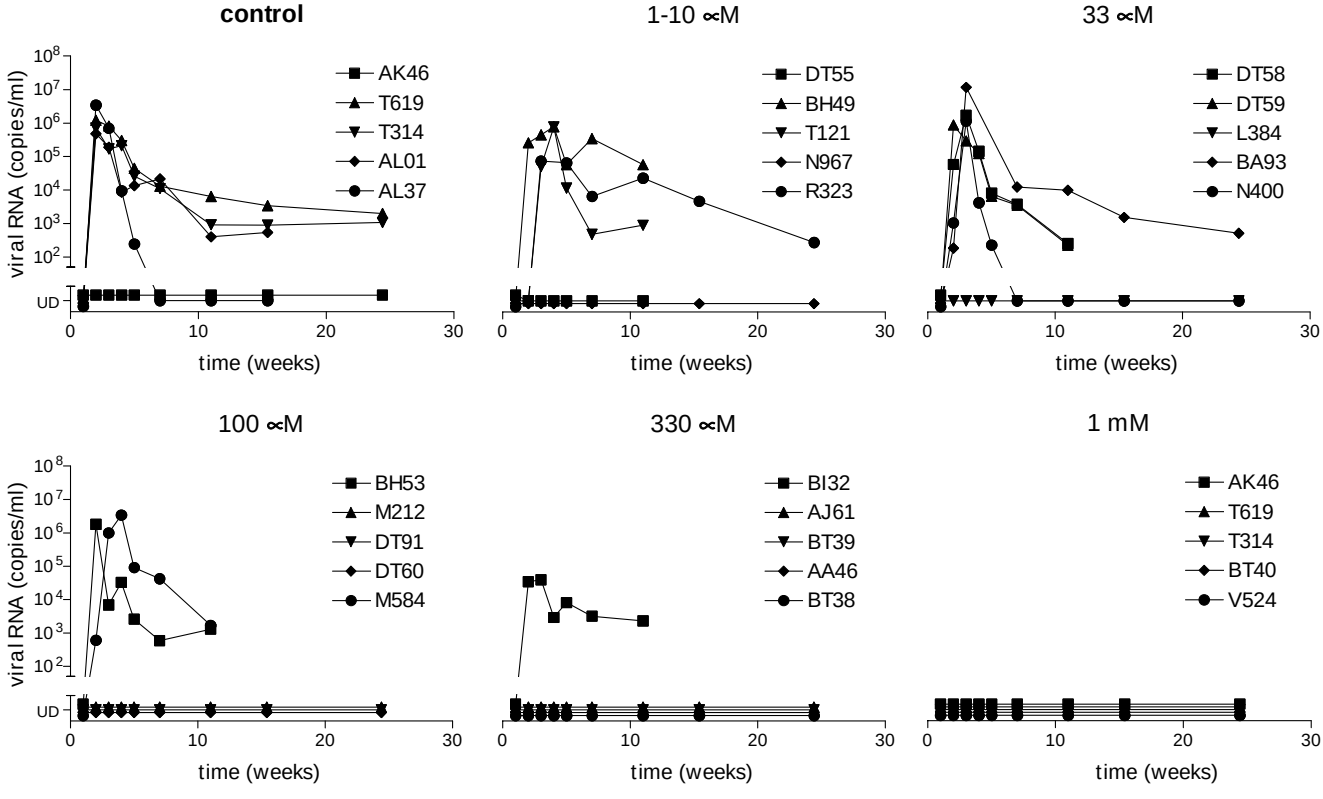


Fig. 3. Topical application of PSC-RANTES blocks infection by SHIV SF162. Six groups of five progesterone-treated rhesus macaques were anesthetized, treated intravaginally with 4 ml PBS control or PSC-RANTES at concentrations from 1 μM to 1 mM in 4 ml PBS, and then challenged 15 min later with 300 $TCID_{50}$ of SHIV SF162. Plasma was monitored as indicated for SHIV RNA by real-time reverse transcription polymerase chain reaction (RT-PCR). UD, undetectable (i.e., below 60 copies per ml).

of our agent against SHIV SF162. Recently, in the same animal model, the neutralizing antibody IgG1-b12 (*31*) gave partial protection against vaginal transmission of SHIV, also at concentrations vastly in excess of those needed in vitro. A small-molecule CCR5 inhibitor, highly potent in vitro, gave only minimal protection in the animal system used here, even as a virtually saturated solution (*32*). Cyanovirin partially inhibited vaginal transmission of a SHIV isolate that also targets CXCR4 but only at concentrations ~10,000 times those required for full inhibition in vitro (*33*). Possible explanations for these dose disparities include incomplete distribution, failure to penetrate to hypothetical submucosal target sites, nonspecific adsorption to mucosal surfaces, or degradation or inhibition by vaginal factors. But conceivably, the explanation might simply be that the progesterone treatment and the dose of SHIV that we and others use to ensure near-universal infection of control macaques [28 of 31 in this system (*28*, *31*, *32*)] constitute an extraordinary challenge. In the natural human setting, risk for acquisition of HIV infection after sexual exposure, although probably not uniform, is on average markedly lower than that shown by unprotected controls in this animal system (*34*). Given the discrepancy between in vitro and in vivo potency, we offer our findings as a proof of principle and as a direction for attempts to render the approach economically acceptable.

PSC-RANTES protected macaques from intravaginal challenge without detectable toxicity or histological changes. Consequently, further development of this and related compounds, either alone or in combination with other agents, and improvement of their formulation are reasonable subjects for further study as an approach to the prevention of sexual transmission of HIV.

References and Notes

1. R. Liu *et al.*, *Cell* **86**, 367 (1996).
2. P. A. Zimmerman *et al.*, *Mol. Med.* **3**, 23 (1997).
3. Y. Huang *et al.*, *Nature Med.* **2**, 1240 (1996).
4. M. Dean *et al.*, *Science* **273**, 1856 (1996).
5. J. R. Salkowitz *et al.*, *Clin. Immunol.* **98**, 200 (2001).
6. E. A. Berger, P. M. Murphy, J. M. Farber, *Annu. Rev. Immunol.* **17**, 657 (1999).
7. P. R. Clapham, A. McKnight, *J. Gen. Virol.* **83**, 1809 (2002).
8. T. B. Geijtenbeek *et al.*, *Cell* **100**, 587 (2000).
9. S. G. Turville *et al.*, *Blood* **98**, 2482 (2001).
10. L. Wu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 1568 (2002).
11. C. J. Miller, R. J. Shattock, *Microbes Infect.* **5**, 59 (2003).
12. M. Pope, A. T. Haase, *Nature Med.* **9**, 847 (2003).
13. C. W. Davis, R. W. Doms, *J. Exp. Med.* **199**, 1037 (2004).
14. G. Simmons *et al.*, *Science* **276**, 276 (1997).
15. D. E. Mosier *et al.*, *J. Virol.* **73**, 3544 (1999).
16. T. Kawamura *et al.*, *J. Exp. Med.* **192**, 1491 (2000).
17. V. S. Torre *et al.*, *J. Virol.* **74**, 4868 (2000).
18. R. Sabbe *et al.*, *J. Virol.* **75**, 661 (2001).
19. O. Hartley *et al.*, *J. Virol.* **77**, 6637 (2003).
20. T. Kawamura *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8401 (2003).
21. D. Mosier, unpublished data.
22. A. Amara *et al.*, *J. Exp. Med.* **186**, 139 (1997).
23. I. Aramori *et al.*, *EMBO J.* **16**, 4606 (1997).
24. G. Alkhatib, M. Locati, P. E. Kennedy, P. M. Murphy, E. A. Berger, *Virology* **234**, 340 (1997).
25. M. Mack *et al.*, *J. Exp. Med.* **187**, 1215 (1998).
26. C. Pastore *et al.*, *Antimicrob. Agents Chemother.* **47**, 509 (2003).
27. Materials and methods are available on *Science* Online.
28. R. Veazey, unpublished data.
29. J. D. Lifson *et al.*, *J. Virol.* **75**, 10187 (2001).
30. O. Hartley, unpublished data.
31. R. S. Veazey *et al.*, *Nature Med.* **9**, 343 (2003).
32. R. S. Veazey *et al.*, *J. Exp. Med.* **198**, 1551 (2003).
33. C. C. Tsai *et al.*, *AIDS Res. Hum. Retroviruses* **20**, 11 (2004).
34. R. H. Gray *et al.*, *Lancet* **357**, 1149 (2001).

35. This work was supported by grants from the NIH (AI 51649 and AI 36219) and from the AIDS Section, Swiss National Science Foundation and was supported in part with U.S. federal funds from the National Cancer Institute, NIH, under contract N01-CO-124000 to J.D.L. and M.P.). We thank S. Cairns, R. Black, J. Kagan, and J. Turpin (National Institute of Allergy and Infectious Diseases) for their support and advice; M. Robertson for administrative assistance; F. Blobel and A. Kung (Gryphon Therapeutics) for provision of PSC-RANTES, and J. Leblanc, K. Medvik, W. Mackay, and M. Dodd for technical assistance. R.E.O. is a cofounder of Gryphon Therapeutics, which focuses on therapeutic applications of synthetic proteins. He both holds equity in the company and is a paid consultant.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/485/DC1
Materials and Methods
References and Notes

19 April 2004; accepted 1 September 2004

Cleavage of proBDNF by tPA/Plasmin Is Essential for Long-Term Hippocampal Plasticity

Petti T. Pang,[1,3] Henry K. Teng,[2] Eugene Zaitsev,[1] Newton T. Woo,[1] Kazuko Sakata,[1] Shushuang Zhen,[2] Kenneth K. Teng,[2] Wing-Ho Yung,[3] Barbara L. Hempstead,[2] Bai Lu[1]*

Long-term memory is thought to be mediated by protein synthesis–dependent, late-phase long-term potentiation (L-LTP). Two secretory proteins, tissue plasminogen activator (tPA) and brain-derived neurotrophic factor (BDNF), have been implicated in this process, but their relationship is unclear. Here we report that tPA, by activating the extracellular protease plasmin, converts the precursor proBDNF to the mature BDNF (mBDNF), and that such conversion is critical for L-LTP expression in mouse hippocampus. Moreover, application of mBDNF is sufficient to rescue L-LTP when protein synthesis is inhibited, which suggests that mBDNF is a key protein synthesis product for L-LTP expression.

Long-lasting changes in synaptic efficacy are thought to mediate long-term memory (*1*, *2*). A well-studied model system is the late phase of long-term potentiation (L-LTP) in the hippocampus. Unlike the early phase of LTP (E-LTP), L-LTP requires new protein synthesis and involves synaptic growth (*2*). A key molecule implicated in L-LTP is the secretory protein BDNF. Hippocampal slices from BDNF heterozygous ($BDNF^{+/-}$) mice fail to exhibit L-LTP (*3*). Inhibition of BDNF signaling by blocking its receptor TrkB with antibody to TrkB or with BDNF scavenger TrkB–immunoglobulin G also inhibits L-LTP (*4*, *5*). In the mammalian brain, BDNF is synthesized as a precursor called proBDNF, which is proteolytically cleaved to generate mature BDNF (mBDNF), the form of BDNF that binds to and activates TrkB (*6*, *7*). In cultured hippocampal neurons and in heterologous cells, proBDNF accounts for a substantial proportion of total BDNF secreted extracellularly (*8*–*11*). A recent study suggests that the precursor and mature forms of neurotrophins interact with very different receptor/signaling systems to induce opposing biological effects (*12*). If a similar principle could be applied to the hippocampus, extracellular cleavage of proBDNF at synapses may have profound implications for synaptic modulation (*7*).

One molecule that may play a role in the conversion of proBDNF to its mature form is the extracellular serine protease tPA. Several studies have implicated tPA in the expression of long-lasting forms of synaptic plasticity: Induction of L-LTP enhances the expression of tPA in the hippocampus (*13*), tPA can be secreted from neuronal growth cones and

[1]Section on Neural Development and Plasticity, Laboratory of Cellular and Synaptic Neurophysiology, National Institute of Child Health and Human Development (NICHD), Bethesda, MD 20892, USA. [2]Division of Hematology, Department of Medicine, Weill Medical College of Cornell University, New York, NY 10021, USA. [3]Department of Physiology, Faculty of Medicine, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

*To whom correspondence should be addressed. E-mail: bailu@mail.nih.gov

axonal terminals (*14*), and neuronal membrane depolarization also induces secretion of tPA into the extracellular space in the hippocampus in a Ca^{2+}-dependent manner (*15*). An inhibitor of tPA blocks L-LTP, whereas extracellular application of tPA results in L-LTP induced by a single tetanus, which otherwise only induces E-LTP (*16*). Mice lacking tPA exhibit a selective deficit in L-LTP expression without affecting E-LTP, and they also show impairment in context conditioning and in two-way active avoidance tests (*17–20*). Transgenic mice overexpressing tPA exhibit enhanced L-LTP and improved spatial learning (*21*).

Although tPA has been implicated in proteolytic degradation of several extracellular matrix proteins (*22–24*), the only well-defined, direct target of tPA is plasminogen. This inactive zymogen is converted to the active form, plasmin, through proteolytic cleavage by tPA (*25*). Plasminogen mRNA and protein in the hippocampus are exclusively expressed in neurons and primarily in the apical dendrites of pyramidal cells (*26*). Because in vitro experiments suggest that plasmin is capable of cleaving proBDNF to its mature form (*12*), we hypothesized that a major function of the tPA/plasmin system is to convert proBDNF to mBDNF at hippocampal synapses, and that such conversion is critical for the expression of L-LTP.

To determine the precise role of BDNF in long-term hippocampal plasticity, we recorded L-LTP in the CA1 area. We used a paradigm consisting of 12 bursts of theta burst stimulation (l-TBS), which is more reliable than the conventional four-tetanus protocol in revealing the L-LTP deficits in $BDNF^{+/-}$ mice (*4*). Similar to previous reports, we consistently observed a severe impairment in L-LTP induced by l-TBS in $BDNF^{+/-}$ mice [Fig. 1A, field excitatory postsynaptic potential (EPSP) slope measured 3 hours after application of tetanus: wild type, 170.1 ± 6.2; $BDNF^{+/-}$, 101.8 ± 8.6; $P < 0.001$].

To investigate whether mBDNF is sufficient for L-LTP, we used two additional approaches. First, we inhibited protein synthesis by a specific inhibitor, anisomycin (40 μM), and showed that L-LTP induced by l-TBS is dependent on protein synthesis. Perfusion of anisomycin throughout the recording effectively blocked L-LTP (Fig. 1B). Application of mBDNF (200 ng/ml) 2 to 3 min after l-TBS completely reversed the blockade of L-LTP by anisomycin (Fig. 1B, anisomycin, 101.3 ± 6.4; anisomycin + mBDNF, 153.4 ± 14.7; $P < 0.05$). Interestingly, the L-LTP deficit could no longer be rescued when mBDNF was applied 10 min after l-TBS, suggesting a critical period in which mBDNF acts (fig. S1). A similar rescuing effect of mBDNF was observed when emetine (20 μM) was used to inhibit protein synthesis (fig. S2). Perfusion of mBDNF to the slices did not alter the basal synaptic transmission over a 3-hour period, nor did it have an epileptic effect on hippocampal neurons (fig. S3). Second, we applied mBDNF to slices that received only three bursts of theta burst stimulation (s-TBS), which normally induces E-LTP but not L-LTP (Fig. 1C). Slices treated with mBDNF after application of s-TBS now exhibited bona fide L-LTP (Fig. 1C, s-TBS, 99.9 ± 10.2; s-TBS + mBDNF, 140.2 ± 10.4; $P < 0.05$).

To study the relationship between tPA and BDNF, we used tPA homozygous ($tPA^{-/-}$) mutant mice. Consistent with previous reports (*17*), we found that there was a severe impairment in L-LTP induced by l-TBS in hippocampal slices derived from the $tPA^{-/-}$

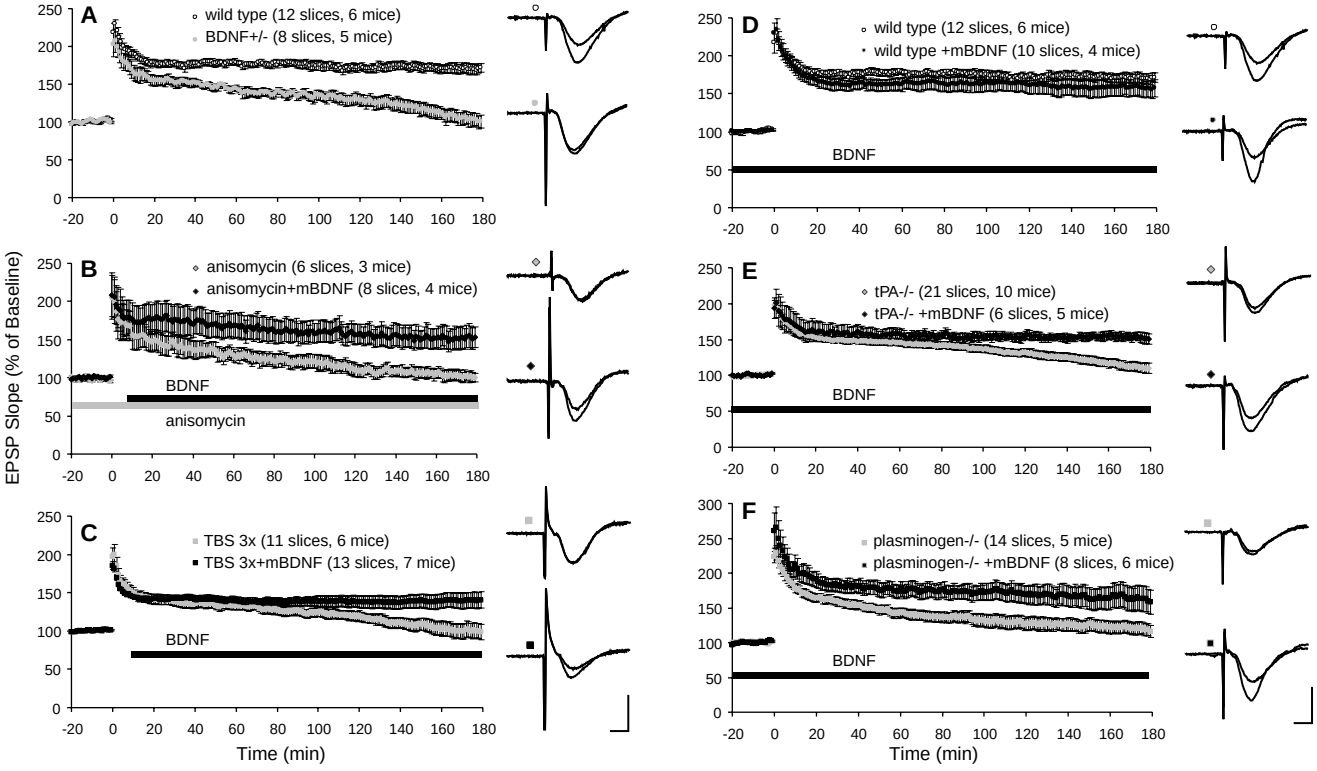


Fig. 1. mBDNF is both necessary and sufficient for L-LTP expression. Field EPSPs were recorded in the CA1 area of hippocampal slices derived from wild-type and various knockout mice. L-LTP was induced by l-TBS; E-LTP was induced by s-TBS. Application of various drugs is indicated by horizontal bars. Numbers of slices and mice used in each condition are indicated at the top of each plot. Superimposed sample traces before and 3 hours after l-TBS are shown at the right of each plot. Scales for these and all other traces: 4 mV (vertical), 4 ms (horizontal). (**A**) BDNF is required for the expression of L-LTP. Note that l-TBS failed to induce L-LTP in $BDNF^{+/-}$ mice. (**B**) mBDNF rescues the impairment of L-LTP caused by inhibition of protein synthesis. Hippocampal slices were treated with the protein synthesis inhibitor anisomycin (40 μM) throughout the entire experiments. mBDNF (200 ng/ml) was applied 5 min after l-TBS, as indicated by the black bar. (**C**) Application of mBDNF after s-TBS converts E-LTP to L-LTP. (**D**) mBDNF does not further enhance L-LTP in wild-type mice. (**E**) mBDNF rescues L-LTP in $tPA^{-/-}$ mice. (**F**) mBDNF rescues L-LTP in $plasminogen^{-/-}$ mice.

mice (Fig. 1E). Perfusion of mBDNF prevented the L-LTP impairment (Fig. 1E, tPA$^{-/-}$, 110.4 ± 6.8; tPA$^{-/-}$ + mBDNF, 151.7 ± 7.4; $P < 0.001$). mBDNF did not simply increase the magnitude of synaptic potentiation, because application of mBDNF to wild-type slices had no effect on L-LTP (Fig. 1D, l-TBS, 170.1 ± 6.2; l-TBS + mBDNF, 156.4 ± 11.5; $P = 0.286$). A direct target of tPA is plasminogen, which is proteolytically cleaved by tPA to form plasmin. We assessed whether plasmin is also involved in L-LTP and, if so, whether a downstream target of plasmin is BDNF. Plasminogen$^{-/-}$ mice also exhibited severe impairment in L-LTP (Fig. 1F). Application of mBDNF completely rescued the L-LTP deficit in these mice (Fig. 1F, plasminogen$^{-/-}$, 118.4 ± 8.6; plasminogen$^{-/-}$ + mBDNF, 161.2 ± 17.8; $P < 0.05$).

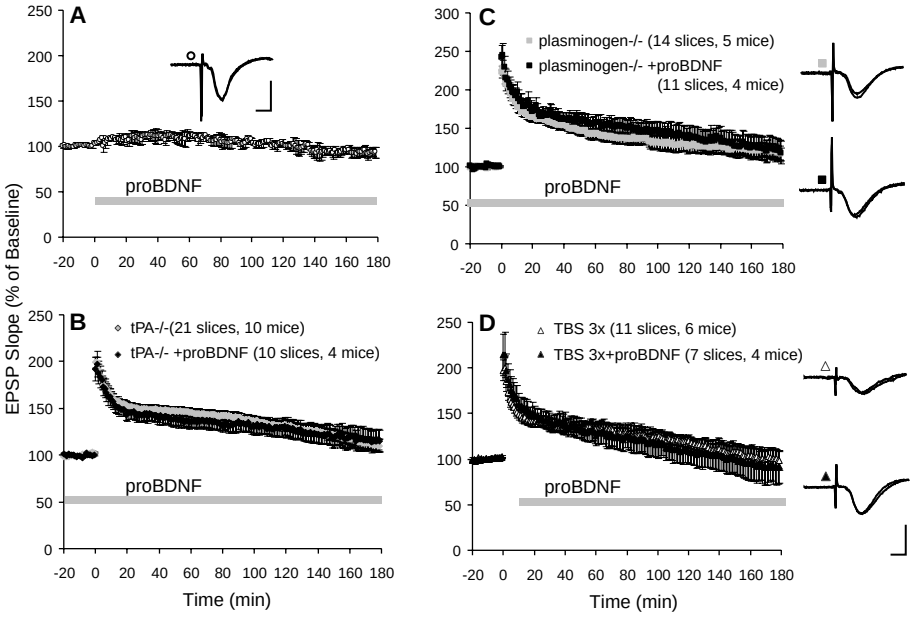


Fig. 2. Cleavage-resistant proBDNF does not mimic the role of mBDNF in L-LTP expression. (**A**) proBDNF has no effect on basal synaptic transmission ($n = 3$ slices; two mice). (**B**) proBDNF fails to rescue L-LTP in tPA$^{-/-}$ mice. (**C**) proBDNF fails to rescue L-LTP in plasminogen$^{-/-}$ mice. (**D**) proBDNF fails to convert E-LTP to L-LTP. s-TBS was applied to CA1 synapses of wild-type hippocampal slices. In (A) and (D), proBDNF (1 to 2 ng/ml) was perfused to the slices as indicated by the horizontal bars; in (B) and (C), proBDNF was applied to the slices for at least 60 min before the delivery of l-TBS.

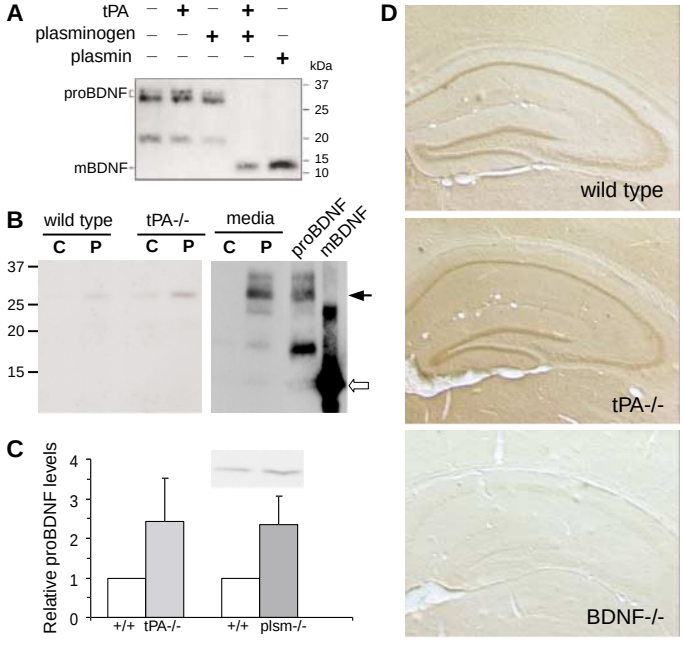


Fig. 3. Plasmin, but not tPA, converts proBDNF to mBDNF. (**A**) Protease cleavage of proBDNF in vitro. Recombinant proBDNF was incubated with tPA, plasmin, plasminogen, or plasminogen plus tPA as indicated. The cleaved products were probed on a Western blot with an antibody to mBDNF. Plasmin and plasminogen plus tPA, but not tPA alone, converted proBDNF to mBDNF. (**B**) Protease cleavage of proBDNF in the hippocampus. Left: Immunoprecipitation analysis shows that, relative to wild-type mice, substantially more proBDNF was detected in hippocampal tissues derived from tPA$^{-/-}$ mice (compare lanes 2 and 4). Right: Antibody specificity. Lysate (lane 7) and media from proBDNF-producing 293T cells (lanes 5 and 6) were immunoprecipitated with either control or specific antibody to proBDNF (366) followed by Western blotting with antibody to mBDNF. Recombinant mBDNF (20 ng) was included as a reference (lane 8). C, PBS coupled; P, proBDNF antibody coupled. (**C**) Summary of relative proBDNF levels in hippocampal tissues derived from tPA$^{-/-}$ and plasminogen$^{-/-}$ mice. (**D**) Protease cleavage of proBDNF in CA1 area. Immunohistochemistry was performed with an antibody specific for proBDNF. More proBDNF immunoreactivity was observed in hippocampal sections derived from tPA$^{-/-}$ mice as compared with those from wild-type mice. No immunoreactivity was detected in sections from BDNF$^{-/-}$ mice.

Plasmin is one of the few secreted extracellular proteases that effectively convert proBDNF to mBDNF through proteolytic cleavage in vitro (*12*). If the cleavage of proBDNF by plasmin was critical for BDNF regulation of L-LTP, a cleavage-resistant proBDNF should fail to rescue the L-LTP deficit seen in plasminogen$^{-/-}$ mice. This was indeed the case. Perfusion of cleavage-resistant proBDNF (1 to 2 ng/ml) was completely ineffective in rescuing the L-LTP deficit in plasminogen$^{-/-}$ mice (Fig. 2C, plasminogen$^{-/-}$, 118.4 ± 8.6; plasminogen$^{-/-}$ + proBDNF, 121.1 ± 14.4; $P = 0.866$). Perfusion of proBDNF also had no effect on basal synaptic transmission (Fig. 2A). Next, we tested the effect of proBDNF on tPA$^{-/-}$ slices. Again, proBDNF failed to rescue L-LTP in tPA mutants that are normally deficient in L-LTP (Fig. 2B, tPA$^{-/-}$, 110.4 ± 6.8; tPA$^{-/-}$ + proBDNF, 115.5 ± 11.4; $P = 0.689$). Finally, unlike mBDNF, application of proBDNF after s-TBS did not allow the expression of L-LTP (Fig. 2D, s-TBS, 99.9 ± 10.2; s-TBS + proBDNF, 91.2 ± 18.5; $P = 0.658$). To ensure that proBDNF at this concentration was biologically active, we performed two different types of assays. First, the same preparation of proBDNF at this concentration elicited a robust apoptosis in smooth muscle cells (*27*). Second, treatment of the slices with proBDNF (1 to 2 ng/ml) enhanced *N*-methyl-D-aspartate (NMDA) receptor–dependent long-term depression (LTD) (*28*).

To determine whether tPA and/or plasmin could directly cleave proBDNF, we performed biochemical analysis in vitro. Purified cleavable proBDNF was incubated with tPA (60 μg/ml), plasmin (20 μg/ml), plasminogen (18 μg/ml), or plasminogen plus tPA. The reaction products were subjected to Western blot with an antibody that detects both proBDNF and mature BDNF. Purified proBDNF exhibited a major doublet band of 30 kD as well as a minor band of 18 kD (Fig. 3A). Neither tPA nor plasminogen alone was capable of cleaving proBDNF in vitro. In contrast, treatment with plasmin converted virtually all proBDNF to mBDNF (Fig. 3A). Although tPA itself was unable to cleave proBDNF, tPA together with plasminogen

was as effective as plasmin in generating mBDNF; this finding suggests that tPA affects proBDNF cleavage indirectly by activating plasmin (Fig. 3A).

Next, we used semiquantitative immunoprecipitation analysis to examine whether the tPA/plasmin system affects proBDNF cleavage in vivo in the hippocampus. Hippocampal tissues from wild-type and $tPA^{-/-}$ mice were dissected. Equal amounts of hippocampal lysates (1.5 mg) were immunoprecipitated with beads coupled to either phosphate-buffered saline (PBS) or an antibody specific for proBDNF, followed by Western blot analysis with an antiserum to mBDNF. Relative to the wild-type control, increased levels of proBDNF in the hippocampi derived from $tPA^{-/-}$ mice were observed (Fig. 3B). Quantitative analysis showed a doubling of the amount of proBDNF in the $tPA^{-/-}$ tissues (Fig. 3C, $n = 3$ pairs of mice). A similar increase in the level of proBDNF was observed in hippocampi derived from $plasminogen^{-/-}$ mice (Fig. 3C, $n = 3$ pairs of mice). Next, to determine whether the proteolytic cleavage occurs in the CA1 region, we performed immunohistochemistry with an antibody specific for proBDNF. Side-by-side comparison revealed greater proBDNF immunoreactivity in the hippocampal sections from $tPA^{-/-}$ mice as compared to those from wild-type mice (Fig. 3D). As a negative control, sections from $BDNF^{-/-}$ mice did not exhibit any immunoreactivity (Fig. 3D). To ensure the reliability of the observation, we performed immunohistochemistry with a second, proBDNF-specific antibody. Again, more immunoreactivity was observed in $tPA^{-/-}$ sections (fig. S4). At a higher magnification, stronger proBDNF staining was seen in the apical dendrites of CA1 pyramidal neurons in $tPA^{-/-}$ mice (fig. S4, right panels). Similar results were obtained with three different pairs of $tPA^{+/+}$ and $tPA^{-/-}$ mice (Fig. 3D) (fig. S4).

Finally, we determined the sequential relationship among tPA, plasmin, and mBDNF in L-LTP expression. The expression and secretion of tPA in hippocampal neurons have been shown to be regulated by neuronal activity (*15*, *16*). Consistent with the extracellular action of tPA, application of tPA (500 ng/ml) to the $tPA^{-/-}$ slices completely reversed the L-LTP deficit (Fig. 4A, $tPA^{-/-}$, 110.4 ± 6.8; $tPA^{-/-}$ + tPA, 166.0 ± 8.1; $P < 0.001$). Exposure of the slices to plasmin for 1 hour (100 nM, from 30 min before to 30 min after l-TBS) also rescued L-LTP deficit in $tPA^{-/-}$ slices (Fig. 4B, $tPA^{-/-}$, 110.4 ± 6.8; $tPA^{-/-}$ + plasmin, 149.7 ± 5.0; $P < 0.001$). Shorter exposure was used to avoid damage of hippocampal neurons by plasmin (*29*). In contrast, application of tPA to $plasminogen^{-/-}$ slices had no effect (Fig. 4C, $plasminogen^{-/-}$, 118.4 ± 8.6; $plasminogen^{-/-}$ + tPA, 103.4 ± 9.4; $P = 0.297$). These results suggest that plasmin is downstream of tPA in controlling the expression of L-LTP. Similarly, mBDNF was able to rescue the L-LTP deficit in $tPA^{-/-}$ slices (Fig. 1E), but tPA failed to do the same in $BDNF^{+/-}$ slices (Fig. 4D, $BDNF^{+/-}$, 101.8 ± 8.6; $BDNF^{+/-}$ + tPA, 95.5 ± 10.9; $P = 0.658$); this result suggests that BDNF is downstream of tPA as well. Moreover, treatment of $plasminogen^{-/-}$ slices with mBDNF produced an excellent late-phase LTP (Fig. 1C), whereas treatment of $BDNF^{+/-}$ slices with plasmin did not rescue the L-LTP deficit (Fig. 4E, $BDNF^{+/-}$, 101.8 ± 8.6; $BDNF^{+/-}$ + plasmin, 108.3 ± 12.4; $P = 0.668$). Taken together, these results support the following model: tPA → plasmin → BDNF in controlling L-LTP.

Although conversion of precursor to mature neurotrophins by extracellular proteases has been shown in vitro (*12*), its physiological role in vivo remains to be established. Our results suggest that such conversion indeed occurs in the brain. We have identified tPA/plasmin as an endogenous extracellular enzyme system, expressed at the hippocampal synapses, that is capable of converting proBDNF to mBDNF. We have shown that tPA, through the activation of plasminogen, converts proBDNF to mBDNF in vitro, and that $tPA^{-/-}$ mice exhibit more proBDNF immunoreactivity in the hippocampus in vivo. Although direct evidence is missing that the tPA/plasmin system cleaves proBDNF extracellularly at the hippocampal synapses, several lines of evidence support this notion. First, most of

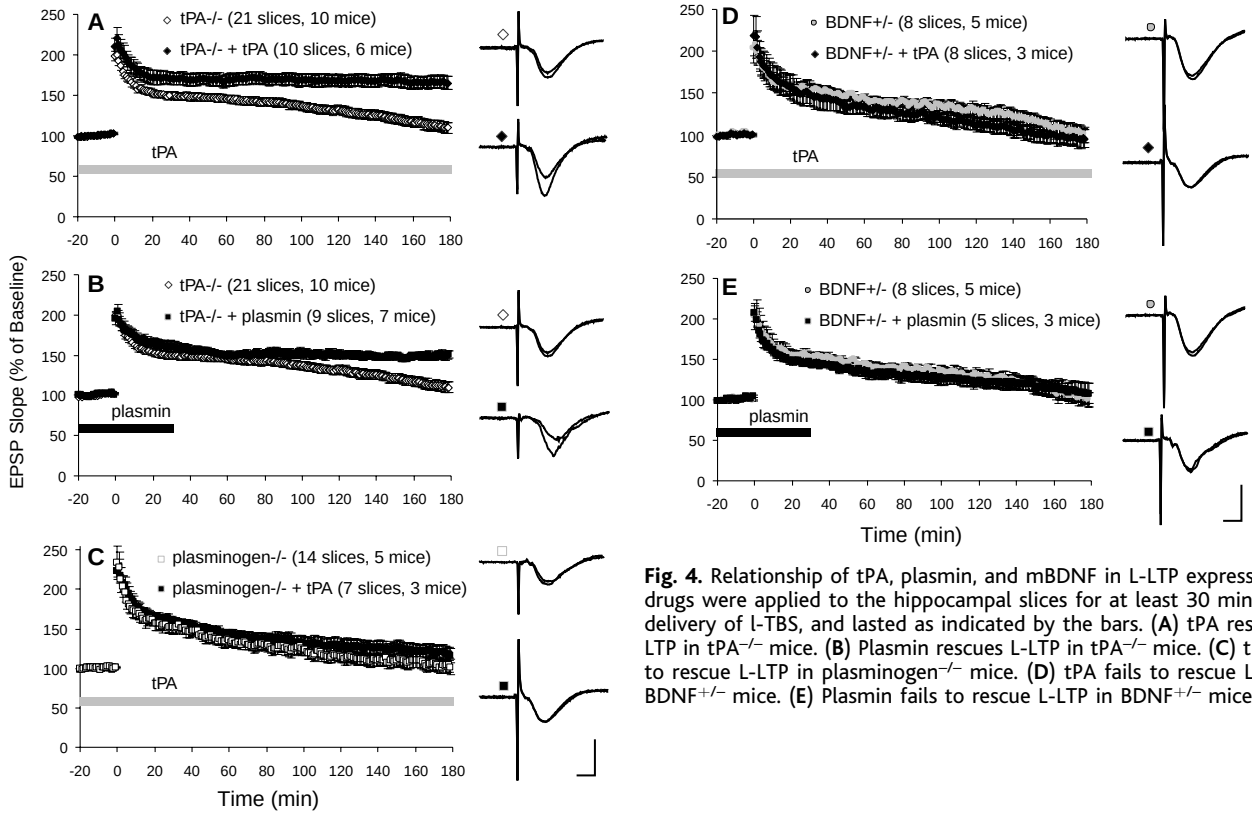


Fig. 4. Relationship of tPA, plasmin, and mBDNF in L-LTP expression. All drugs were applied to the hippocampal slices for at least 30 min before delivery of l-TBS, and lasted as indicated by the bars. (**A**) tPA rescues L-LTP in $tPA^{-/-}$ mice. (**B**) Plasmin rescues L-LTP in $tPA^{-/-}$ mice. (**C**) tPA fails to rescue L-LTP in $plasminogen^{-/-}$ mice. (**D**) tPA fails to rescue L-LTP in $BDNF^{+/-}$ mice. (**E**) Plasmin fails to rescue L-LTP in $BDNF^{+/-}$ mice.

the BDNF secreted by hippocampal neurons appears to be in the precursor form (*8*–*11*). Second, the secretion of proBDNF is activity dependent (*9*). Moreover, both tPA and plasmin are expressed and secreted at hippocampal synapses (*15*, *16*, *26*). It is conceivable that application of l-TBS induces the secretion of proBDNF or the activation of the tPA/plasmin system (or both) at the CA1 synapses, and that subsequent conversion of proBDNF to mBDNF plays an essential role in L-LTP expression.

BDNF and tPA are perhaps the two best-characterized secretory molecules involved in L-LTP and long-term memory. Our study provides a mechanistic link between these two seemingly independent molecule systems in L-LTP expression. We have also worked out the sequential relationship among tPA, plasmin, and mBDNF. Thus, the extracellular protease tPA cleaves plasminogen to generate plasmin, which in turn converts the precursor proBDNF to mBDNF. A new mechanism that regulates late-phase LTP could be the activity-dependent extracellular cleavage of proBDNF. Given that proBDNF preferentially activates pan neurotrophin receptor p75NTR over TrkB receptor, it is also possible that the inhibition of proteolytic conversion of proBDNF may lead to a negative regulation of hippocampal plasticity.

mBDNF applied after the delivery of l-TBS appears to be sufficient to maintain L-LTP in slices incubated in the protein synthesis inhibitor anisomycin (40 μM) during the entire course of recording (Fig. 1B). At this concentration of anisomycin, all protein synthesis is completely blocked. Consistent with this finding, perfusion of mBDNF after the application of the E-LTP–inducing s-TBS converts E-LTP to L-LTP (Fig. 1C).

It has long been believed that L-LTP and long-term memory requires new protein synthesis, but the specific product(s) mediating the long-term changes is not known. Our results imply that mBDNF is a key protein synthesis product, if not the only one, needed to carry on all the necessary functions for long-term modification of hippocampal synapses. L-LTP involves long-lasting changes in the structure and function of the hippocampal synapses. BDNF elicits long-lasting enhancement of synaptic transmission (*30*, *31*), promotes dendritic arborization (*32*), and stimulates the growth of dendritic spines (*33*). Activity-dependent BDNF synthesis has been shown with the use of L-LTP–inducing tetanus (*34*). The key function of the L-LTP–inducing tetanus could be to induce the synthesis and/or processing of proBDNF, and the newly generated mBDNF would then be responsible for eliciting all the structural and functional changes underlying L-LTP at the CA1 synapses.

References and Notes

1. T. Abel *et al.*, *Cell* **88**, 615 (1997).
2. E. R. Kandel, *Science* **294**, 1030 (2001).
3. M. Korte, H. Kang, T. Bonhoeffer, E. Schuman, *Neuropharmacology* **37**, 553 (1998).
4. S. L. Patterson *et al.*, *Neuron* **32**, 123 (2001).
5. H. Kang, A. A. Welcher, D. Shelton, E. M. Schuman, *Neuron* **19**, 653 (1997).
6. T. Li *et al.*, *Biopolymers* **67**, 10 (2002).
7. B. Lu, *Neuron* **39**, 735 (2003).
8. M. F. Egan *et al.*, *Cell* **112**, 257 (2003).
9. Z.-Y. Chen *et al.*, *J. Neurosci.* **24**, 4401 (2004).
10. S. J. Mowla *et al.*, *J. Biol. Chem.* **276**, 12660 (2001).
11. S. J. Mowla *et al.*, *J. Neurosci.* **19**, 2069 (1999).
12. R. Lee, P. Kermani, K. K. Teng, B. L. Hempstead, *Science* **294**, 1945 (2001).
13. Z. Qian, M. E. Gilbert, M. A. Colicos, E. R. Kandel, D. Kuhl, *Nature* **361**, 453 (1993).
14. A. Krystosek, N. W. Seeds, *Science* **213**, 1532 (1981).
15. A. Gualandris, T. E. Jones, S. Strickland, S. E. Tsirka, *J. Neurosci.* **16**, 2220 (1996).
16. D. Baranes *et al.*, *Neuron* **21**, 813 (1998).
17. Y. Y. Huang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8699 (1996).
18. U. Frey, M. Muller, D. Kuhl, *J. Neurosci.* **16**, 2057 (1996).
19. P. Calabresi *et al.*, *Eur. J. Neurosci.* **12**, 1002 (2000).
20. R. Pawlak *et al.*, *Neuroscience* **113**, 995 (2002).
21. R. Madani *et al.*, *EMBO J.* **18**, 3007 (1999).
22. K. B. Hoffman, J. Martinez, G. Lynch, *Brain Res.* **811**, 29 (1998).
23. Y. P. Wu *et al.*, *J. Cell Biol.* **148**, 1295 (2000).
24. Y. Nakagami, K. Abe, N. Nishiyama, N. Matsuki, *J. Neurosci.* **20**, 2003 (2000).
25. E. F. Plow, T. Herren, A. Redlitz, L. A. Miles, J. L. Hoover-Plow, *FASEB J.* **9**, 939 (1995).
26. S. E. Tsirka, A. D. Rogove, T. H. Bugge, J. L. Degen, S. Strickland, *J. Neurosci.* **17**, 543 (1997).
27. The ability of proBDNF to induce apoptosis in p75-expressing vascular smooth muscle cells was assessed as described (*12*) using terminal deoxynucleotidyl transferase–mediated deoxyuridine triphosphate nick end labeling (TUNEL) detection. Treatment of cells with proBDNF (2 ng/ml) resulted in 11.2 ± 1.8% TUNEL$^+$ cells, whereas the vehicle-treated cells exhibited only 2.1 ± 1.1% TUNEL$^+$ cells.
28. Although application of proBDNF failed to affect basal transmission and L-LTP, it facilitated NMDA receptor–dependent LTD in the hippocampus. In 8-week-old, wild-type hippocampal slices, LTD induced by low-frequency stimulation (LFS; 1 Hz, 15 min) was significantly higher (29.5 ± 2.4%, $n = 11$) in slices treated with proBDNF as compared to untreated slices (8.0 ± 6.6%, $n = 10$) ($P < 0.01$).
29. Plasmin is known to degrade a variety of substrates, including the extracellular matrix proteins laminin, fibrin, and fibronectin; such degradation leads to cell death. In our own work, we found that plasmin caused detrimental effects to L-LTP in wild-type slices if treated more than 2 hours. In contrast, a short-term exposure was without effect. Thus, to avoid the damaging effect of long-term plasmin exposure, we used a protocol that treated the slices for only 60 min. As indicated in Fig. 4C, this short-term treatment was sufficient to rescue the L-LTP defects.
30. H. Kang, E. M. Schuman, *Science* **267**, 1658 (1995).
31. L. C. Rutherford, S. B. Nelson, G. G. Turrigiano, *Neuron* **21**, 521 (1998).
32. A. K. McAllister, D. C. Lo, L. C. Katz, *Neuron* **15**, 791 (1995).
33. W. J. Tyler, L. D. Pozzo-Miller, *J. Neurosci.* **21**, 4249 (2001).
34. S. Patterson, L. M. Grover, P. A. Schwartzkroin, M. Bothwell, *Neuron* **9**, 1081 (1992).

35. We thank R. Desimone and members of the Lu laboratory for the thoughtful comments and suggestions, and Regeneron Pharmaceuticals for providing recombinant BDNF. Supported by the NICHD intramural research program (B.L.) and NIH grant NS30658 (B.L.H.). Molecular interaction data have been deposited in the Biomolecular Interaction Network Database with accession code 153566.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5695/487/DC1
Materials and Methods
Figs. S1 to S4
References

10 May 2004; accepted 26 August 2004

Molecular Architecture of the KvAP Voltage-Dependent K$^+$ Channel in a Lipid Bilayer

Luis G. Cuello, D. Marien Cortes, Eduardo Perozo*

We have analyzed the local structure and dynamics of the prokaryotic voltage-dependent K$^+$ channel (KvAP) at 0 millivolts, using site-directed spin labeling and electron paramagnetic resonance spectroscopy. We show that the S4 segment is located at the protein/lipid interface, with most of its charges protected from the lipid environment. Structurally, S4 is highly dynamic and is separated into two short helices by a flexible linker. Accessibility and dynamics data indicate that the S1 segment is surrounded by other parts of the protein. We propose that S1 is at the contact interface between the voltage-sensing and pore domains. These results establish the general principles of voltage-dependent channel structure in a biological membrane.

Voltage-dependent channels are composed of two functionally linked but structurally independent domains (*1*–*4*). The pore domain is responsible for ion selectivity and contains the channel gate, whereas a voltage-sensing domain (segments S1 to S4) alters the conformation of the gate in response to changes in transmembrane voltage. Crystal structures of

three different prokaryotic K^+ channels have elegantly demonstrated the common architecture of the pore domain (*5*–*7*), as well as the basic principles underlying permeation and selectivity for K^+ ions (*8*, *9*). Furthermore, they point to at least one plausible mechanism for the opening of the intracellular gate (*10*, *11*). There is much less agreement in relation to the structure and conformations of the voltage-sensor domain.

Voltage sensing is linked to structural rearrangements of the S4 segment, a transmembrane helix containing positive charges every three residues. Based on functional and indirect structural analyses, the general consensus has been that the S4 segment and its charges must be isolated from the low dielectric of the membrane by a shield of protein. In one explicit model, this shield surrounds the S4 segment, which moves across an aqueous "gating pore" or "canaliculi" in response to changes in the electric field (*12*–*14*). However, structural and mechanistic models derived from the recent crystal structures of KvAP and its isolated voltage sensor (*15*) appear to run contrary to these concepts. Obtained as complexes with an Fab antibody fragment, the new structures led to the suggestion that S4 and parts of S3 form a stable hairpin (the "paddle") that is located at the periphery of the channel, exposed to the membrane lipid (*15*). This multicharged hairpin would act as a hydrophobic cation moving across the membrane and pulling on the activation gate, thus opening the channel (*16*).

To evaluate these seemingly incompatible models, we have used site-directed spin labeling and electron paramagnetic resonance (EPR) spectroscopy (*17*–*19*) to measure the structural dynamics of reconstituted KvAP. Experiments were carried out on a set of mutants comprising the entire voltage-sensing domain (*20*) under conditions that promote a deep inactivated state, in which the voltage sensor is presumably in a conformation similar to that in the open state. Our data were evaluated within the framework of four distinct structural references: the KvAP crystal structure (*15*), the structure of the KvAP isolated voltage sensor (*15*), a structural model (*21*) based on the open KvAP paddle model (*16*), and models with "canonical" voltage-sensor arrangements (*22*, *23*).

Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22906, USA.

*To whom correspondence should be addressed. E-mail: eperozo@virginia.edu

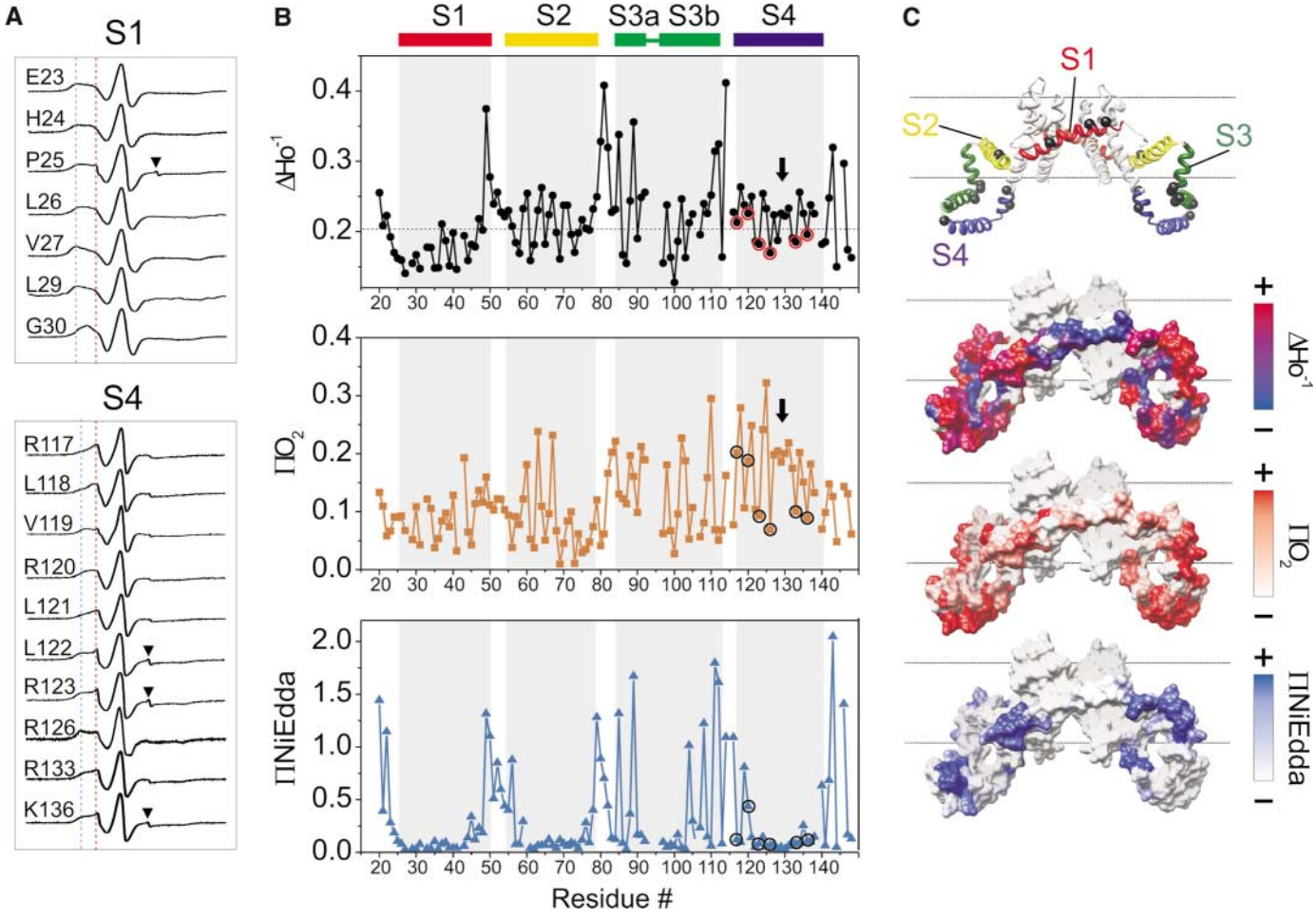


Fig. 1. Voltage-sensor environmental data set. (**A**) EPR spectra of spin-labeled mutants from selected regions of S1 and S4 segments. Spectra represent 100 G scans. Dotted lines mark the location of restricted (blue) and highly dynamic motional components of the spectra. Arrowheads point to the small amounts of residual free spin label. (**B**) Environmental parameter profiles: mobility parameter ΔH_0^{-1} (black circles), oxygen accessibility parameter ΠO_2 (red squares), and NiEdda accessibility parameter ΠNiEdda (blue triangles). Grayed areas represent assignments derived from the KvAP crystal structure. The dotted line is the average mobility for the entire segment. Arrows in the S4 segment point to residues that display a loss in α-helical periodicity. Gating charges in S4 are highlighted by circles. (**C**) Environmental parameters (ΔH_0^{-1}, ΠO_2, and ΠNiEdda) mapped onto a molecular surface rendering of the full-length KvAP crystal structure. (Top) Ribbon representation of the full-length KvAP crystal structure (two subunits are shown for clarity). Individual transmembrane segments are color coded as follows: S1, red; S2, yellow; S3, green; and S4, blue. Black spheres show the location of cysteine mutants for which we were unable to obtain data. Molecular surface and data mapping was done using Chimera (*35*, *36*).

The fully processed EPR data set for the whole sensor domain is shown in Fig. 1 (*24*). Figure 1A shows two representative sets of spectra illustrating the overall distribution of line shapes for regions of S1 and S4 (including all charged positions). The spectra revealed populations of mixed dynamics (particularly in the less mobile positions in S4), suggesting that the sensor as a whole is highly flexible when compared with the overall rigidity of the pore domain as observed in KcsA (*25*, *26*). Figure 1B shows the residue environmental parameter profiles for probe mobility ΔH_0^{-1} (black circles), O_2 accessibility ΠO_2 (red squares), and Ni^{++} chelate complex (NiEdda) accessibility Π_{NiEdda} (blue triangles) from reconstituted full-length KvAP. We found clear, well-defined boundaries to the individual transmembrane (TM) segments, both in terms of probe dynamics and exposure to the aqueous environment in the connecting loops (Fig. 1B, top and bottom). Loops tend to be highly mobile and readily accessible to NiEdda, suggesting an unambiguous transmembrane orientation for all TM segments. This topological arrangement is incompatible with the disposition of both S1 and S2 segments in the crystal structure (Fig. 1C), as well as with that proposed for S2 in the structural model of open KvAP (*15*, *16*).

In the bilayer-embedded regions, average motional dynamics gradually increases from S1 through S3 to S4, with a somewhat parallel increase in O_2 accessibility, in which S4 clearly emerges as the TM segment with the largest O_2 accessibility. In S4, the α-helical periodicity in ΔH_0^{-1} and ΠO_2 is interrupted around residues 128 to 130, dividing the segment into two smaller helices (S4a and S4b) joined by a somewhat flexible linker (Fig. 1B, black arrows). In addition, we find that a large number of residues in S3 are exposed to collisions with NiEdda (Fig. 1B, bottom), although they are well within the confines of the putative transmembrane region, according to the structure of the isolated sensor.

Although there is widespread agreement that the full-length KvAP structure is distorted by the crystallization conditions, mapping the EPR-determined environmental parameters directly onto it provides a structural context to evaluate the degree of distortion (Fig. 1C). Local dynamics do not reveal many gross discrepancies with our experimental results (with the exception of S1), yet it is clear that among the regions most exposed to O_2 (S3 and S4 and parts of S2), few appear to be within the expected bilayer boundaries, according to the location of the pore domain. The situation is precisely the opposite for positions exposed to NiEdda, where the loops and portions of S3 appear almost in the center of the bilayer.

Mapping the same data set onto the structure of the isolated sensor domain (Fig. 2) revealed a better spatial correlation, which suggests that the structure of the isolated sensor is a reasonable representation of its conformation as part of the full-length channel. The water/lipid interface is well defined from the NiEdda map and comprises crescent-shaped surfaces in both ends of the sensor domain, along its long axis. As expected, these are populated by residues from the N terminus, the S1-S2 loop, the S2-S3 loop, and S3b and the extended C-terminal end of S4. Mapped O_2 and mobility data also help position the sensor relative to the lipid and the pore domain. S1 and portions of S2 outline a motionally restricted, low O_2 accessibility area that represents the likely docking surface of the sensor to the pore domain. At the opposite side, S4 and S3a form an X-shaped, highly dynamic, and O_2-accessible area expected to be at the protein/lipid interface.

A more detailed investigation of the likely arrangement of the individual TM segments can be obtained from frequency and vector analysis of the residue accessibility parameters (Fig. 3). In each case (Fig. 3, A to C), the resultant vectors displaying the orientation of the accessible surfaces are shown both in the context of helical wheel representations and mapped in three dimensions on the individual segment structures derived from the isolated sensor. S1 and S2 offer contrasting examples of this analysis (Fig. 3A). In S1, the sum vectors for O_2 accessibility ($\overrightarrow{\Pi O_2}$) and mobility ($\overrightarrow{\Delta H_O^{-1}}$) are negligible, which strongly suggests that S1 is likely surrounded by other regions of the channel. In S2, $\overrightarrow{\Pi O_2}$ and $\overrightarrow{\Delta H_O^{-1}}$ are robust and essentially in phase, pointing to a well-defined lipid-exposed region toward the center of the helix. The very large sum vectors in S3a and S3b for both Π_{NiEdda} and ΔH_0^{-1} are also in phase (Fig. 3B), which is consistent with the idea that the putative aqueous crevices are large enough to allow substantial motions of the spin label. Indeed, the NiEdda accessibility profile shows that a stretch of only about 14 residues in S3 is fully inaccessible to NiEdda, corresponding to a slab of about 20 Å or less across the membrane. Moreover, because the Π_{NiEdda} parameter estimates effective collisions for a probe of 5 to 6 Å, it is likely that actual water penetration into these putative crevices is more extensive than that revealed by the Π_{NiEdda} parameter.

Based on the loss of α-helical periodicity around residue 129, the calculated $\overrightarrow{\Pi O_2}$ and ΔH_0^{-1} for both S4a and S4b are large and essentially in phase (Fig. 3C), as expected from an S4 with an extensive lipid-accessible surface. Additional data about the peripheral location of S4, based on fluorescence measurements of aggregated channel, are provided in (*27*). Given the overwhelming evidence for the peripheral location of S4, a key question to ask is what type of environment surrounds the charged residues. Most of the S4 charges in KvAP (R123, R126, R133, and R136) are not exposed to the environment, as indicated by low probe mobility and negligible accessibility to either O_2 or NiEdda (Fig. 3C, blue numerals). R120 appears partially protected, although it is at the edge of the nonaccessible region in the helical

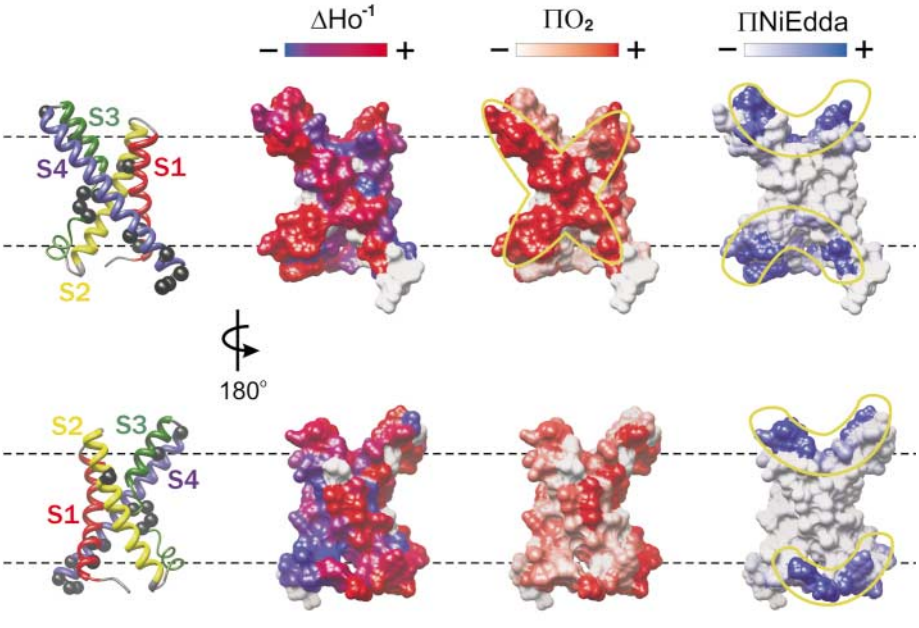


Fig. 2. EPR data mapped onto the isolated voltage-sensor structure. Yellow outlines demarcate the regions of the sensor with the largest accessibility to O_2 or NiEdda. The dashed black lines point to the presumed limits of the lipid bilayer. Details are identical to those in Fig. 1C.

wheel (Fig. 3C). Finally, R117 is located in the accessible region of the helical wheel and shows fairly high motional freedom, but its intermediate O_2 and NiEdda accessibility argues for a location close to the water/lipid interface. In the presence of the S4a-S4b linker, the two resulting helices appear to be twisted ~90° in relation to their accessible surface, an arrangement that places the majority of the charges on the same face of the segment, in contrast with the natural helical screw arrangement of charges expected if the S4 were a straight helix (Fig. 3D).

The present data set provides an opportunity to evaluate explicit open-state models of potassium channel structure. To this end, we have considered two recent versions of the "canonical" model from the laboratories of Benoit Roux and Diane Papazian [the LPR model (*23*)] and Robert Guy [the DHG model (*22*)] (Fig. 4A, data shown for the LPR model only), plus a paddle-like structural model (*21*) based on an interpretation of the KvAP open "structure" of Jiang *et al.* (Fig. 4D) (*15*). [Details on the construction of the model are given in (*27*)]. Therefore, these coordinates are an approximation of the original paddle model (which does not include an explicit position for S1) and thus should serve only as a guide for the general evaluation of the model in relation to the present data set. When considering the water-exposed surfaces, the canonical models show remarkable compatibility with the ΠNiEdda experimental data, as expected from the transmembrane topology of all helical segments (Fig. 4B). However, the agreement is not favorable for the paddle model (Fig. 4E), which places extensive areas of NiEdda-accessible residues well within the low dielectric regions of the bilayer. The bulk of these discrepancies derive from the unusual placement of S2 as a band surrounding the pore domain, approximately parallel to the plane of the bilayer. The lipid-exposed regions were evaluated more quantitatively by correlating the average of the experimentally determined O_2 accessibility in each TM segment with the solvent accessibility calculated by a hard-sphere scanning method (Fig. 4, C and F) (*28*). In this case, the canonical models fail to produce a positive correlation with the calculated accessibilities, primarily as a result of the shielding of S4 away from the periphery of the molecule (Fig. 4C). On the other hand, the paddle model shows significant correlation to the expected TM segment accessibilities, a consequence of the location of the paddle at the channel/lipid interface (Fig. 4F).

We found that many of these discrepancies can be reduced or eliminated by reasonably simple reorientations of the sensor domain structure relative to the pore domain. In the case of both canonical models, a rotation of ~100° to 120° degrees about the long axis of the sensor domain helps correct major discrepancies in lipid-exposed accessibilities. In the paddle model, tilting the sensor domain ~50° to 60° toward the pore domain (plus internal repositioning of S1 and S2) places the majority of the water-exposed regions above the water/lipid interface, preserving the disposition of the lipid-exposed areas. This first-order approximation simply docks the current isolated sensor structure onto the pore domain, although it is likely that the sensor structure will be somewhat different in the context of the full-length channel. The resulting model (Fig. 4G) places the S4 segment squarely at the protein/lipid interface, in agreement with the paddle model. The voltage sensor contacts the pore domain in the immediate vicinity of the S1 segment, whose restricted dynamics and low accessibility imply that it is mostly surrounded by protein. Furthermore, the S4 segment behaves as two α helices, S4a and S4b (Fig. 3), connected by a short linker (residues 129 to 131). The influence of this linker region on the functional behavior of the sensor remains to be established, but it might point to the presence of a hinge that leads to differential rearrangements of these two regions of S4 in response to voltage changes.

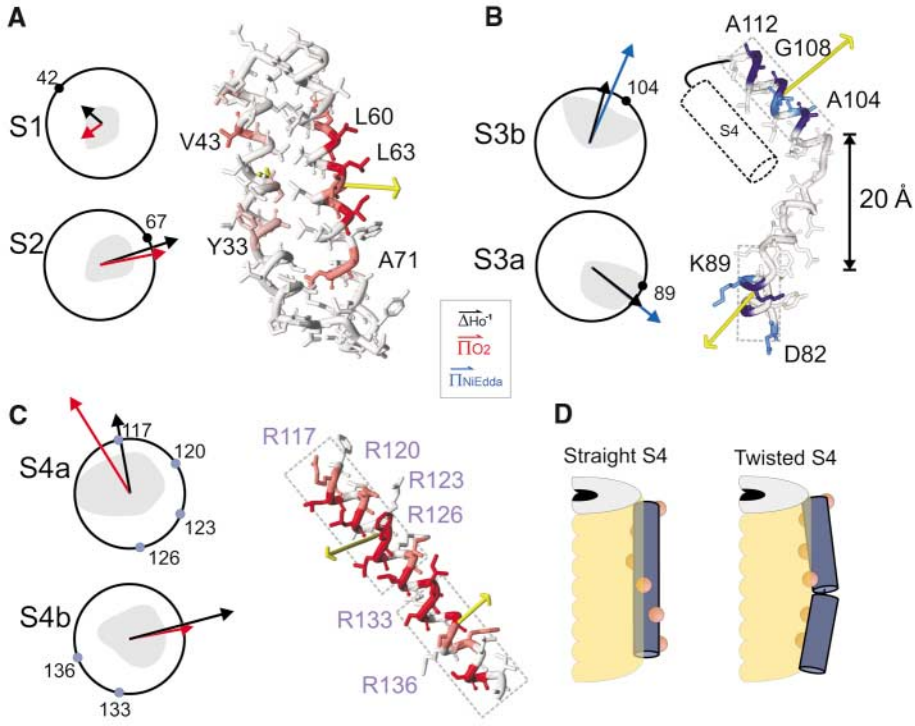


Fig. 3. Frequency and vector analysis of environmental data point to the likely arrangement of the individual TM segments. (**A** to **C**) The orientation of the sum vector for accessibility $\overrightarrow{\Pi O_2}$ and mobility data $\overrightarrow{\Delta H_o^{-1}}$ in S1-S2 (**A**), S3 (**B**), and S4 (**C**) as shown on helical projections (circle), with a single residue serving as reference point (black dot). In S4, the orientation of the charged residues is represented by blue dots. The shaded area highlights the degree of eccentricity for the complete set of accessibility data relative to the maximal accessibility vector. In each case, the sum vector pointing to the direction of highest accessibility is also plotted in three dimensions (yellow arrows) in relation to the structure of each segment as it appears in the isolated voltage sensor. Accessibility to O_2 or NiEdda has been color coded onto the backbone worm of each TM segment. In segments S3 and S4, helical wheels are also shown for the two individual helices between the linker (S3a, S3b, S4a, and S4b) and the relevant regions highlighted by a gray dotted rectangle around the structure. In the S4 segment, the position of the charged residues is shown by blue dots around the unitary circle. (**D**) Conceptual model illustrating the effect of twisting S4a and S4b (relative to each other) on the location of charged groups at the protein-membrane interface.

Our data support the general idea of the S4 charges being shielded from the low dielectric environment of the membrane (at least for the open/inactivated conformation), in agreement with general concepts behind earlier voltage-sensor models. However, the present results are incompatible with models that place the S4 segment in a cocoon of surrounding protein that protects it from the lipid environment and generates a so-called "S4 channel" or "canaliculi." The specific pattern of dynamics and lipid accessibilities for S1 to S3 also makes our data incompatible with more recent transitional models in which the S4 segment might be partially exposed to lipid (*22*, *29*). At the moment, we do not have enough constraints to differentiate between a paddle-like model

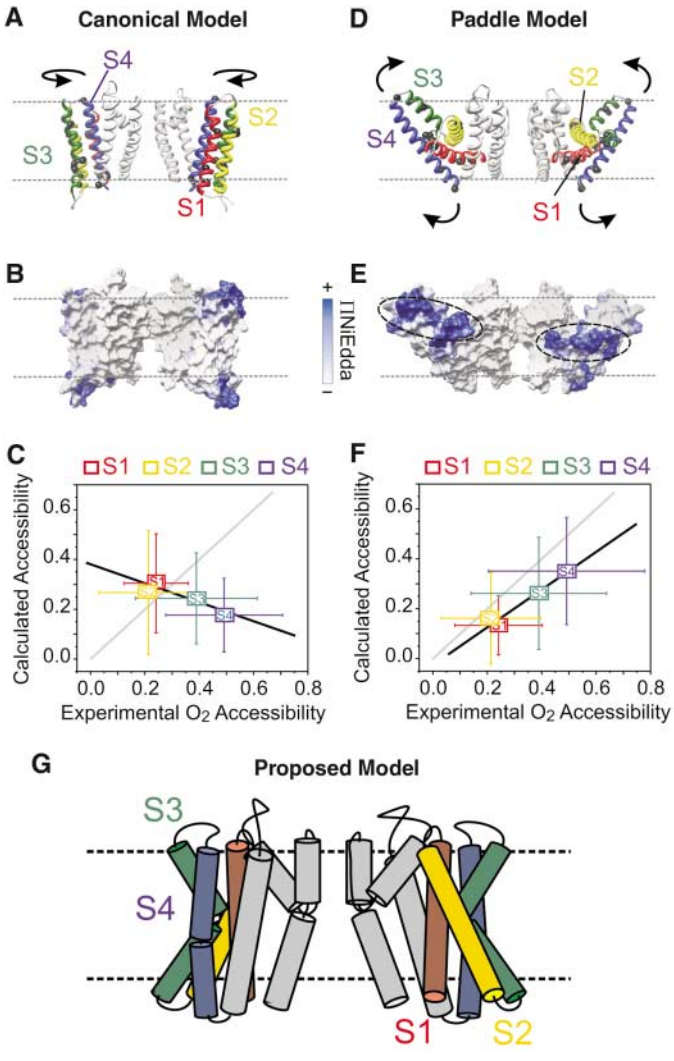


Fig. 4. Evaluation of structural models of KvAP. The explicit coordinates of three structural models of open KvAP were used to evaluate overall concordance with experimental accessibility data. (**A**) Ribbon representation of the canonical model (results for the Laine-Papazian-Roux model are shown). Color coding of the TM segments is the same as in Fig. 2C. (**B**) Map of NiEdda-accessible residues onto a solvent accessible surface. The dotted line points to the approximate location of the water/lipid interface. Black ovals highlight regions of high NiEdda accessibility putatively embedded in the bilayer. (**C**) Correlation between the average solvent accessibility (*28*) of the lipid-embedded region of the LPR model, with the average experimentally determined O_2 accessibility in the same set of residues. Each point corresponds to individual TM segments; the error bars represent SDs. The gray line represents a 1:1 correlation. (**D**) Ribbon representation of the paddle model, as above. (**E**) Map of NiEdda-accessible residues onto a solvent accessible surface. (**F**) Correlation between the average solvent accessibility of the lipid-embedded region of the paddle model with the average experimentally determined O_2 accessibility in the same set of residues. Details as in (C). In parts (B) and (E), the solvent accessible surface was calculated and color mapped with the program Chimera (*35*, *36*). (**G**) Side view of a helix-packing model of KvAP based on the present data set showing the arrangement of the TM segments relative to the membrane and the pore domain. The model highlights the peripheral location of the S4 segment and points to S1 as the segment most likely to be surrounded by protein.

and one in which the S4 segment is indeed peripheral but is also part of deep aqueous crevices, helping to "focus" the transmembrane voltage field during channel activation (*30*, *31*). This will require information regarding the conformation of the sensor in the closed state.

Our model is also partially at odds with the results from studies of perturbation analyses in eukaryotic Kv channels (*32*–*34*). Although there is agreement regarding the local environment surrounding S2 and S3, these studies have concluded that S1 is likely located at the periphery of the channel, in contrast with the present results. We do not fully understand the origin of this discrepancy, although the direct translation of functional data into structural parameters is not always straightforward. Although further structural analyses will be required to solve these issues and ultimately define the multiple conformations of membrane-embedded Kv channels, the present KvAP model represents a starting point for the analysis of the structure and conformational changes underlying voltage-dependent gating.

References and Notes

1. L. Y. Jan, Y. N. Jan, *J. Physiol.* **505**, 267 (1997).
2. F. Bezanilla, *Physiol. Rev.* **80**, 555 (2000).
3. W. A. Catterall, *Neuron* **26**, 13 (2000).
4. G. Yellen, *Nature* **419**, 35 (2002).
5. D. A. Doyle *et al.*, *Science* **280**, 69 (1998).
6. Y. Jiang *et al.*, *Nature* **417**, 515 (2002).
7. A. Kuo *et al.*, *Science* **300**, 1922 (2003).
8. Y. Zhou, J. H. Morais-Cabral, A. Kaufman, R. MacKinnon, *Nature* **414**, 43 (2001).
9. S. Berneche, B. Roux, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8644 (2003).
10. Y. Jiang *et al.*, *Nature* **417**, 523 (2002).
11. T. Jin *et al.*, *Mol. Cell* **10**, 469 (2002).
12. N. Yang, A. L. George Jr., R. Horn, *Neuron* **16**, 113 (1996).
13. H. P. Larsson, O. S. Baker, D. S. Dhillon, E. Y. Isacoff, *Neuron* **16**, 387 (1996) .
14. S. A. Goldstein, *Neuron* **16**, 717 (1996) .
15. Y. Jiang *et al.*, *Nature* **423**, 33 (2003).
16. Y. Jiang, V. Ruta, J. Chen, A. Lee, R. MacKinnon, *Nature* **423**, 42 (2003).
17. W. L. Hubbell, A. Gross, R. Langen, M. A. Lietzow, *Curr. Opin. Struct. Biol.* **8**, 649 (1998).
18. H. Mchaourab, E. Perozo, in *Distance Measurements in Biological EPR*, G. Eaton, S. Eaton, L. Berliner, Eds. (Kluwer, New York, 2000).
19. H. J. Steinhoff, B. Suess, *Methods* **29**, 188 (2003).
20. Starting with a cysteine-less construct of full-length KvAP, we generated 144 cysteine mutants across the entire sequence of the voltage-sensor domain (S1 to S4) and the C-terminal end of S6. Of these, 131 mutants yielded protein suitable for structural studies (>90% coverage), as determined from its oligomeric state in gel-filtration analysis (not shown). Individual mutants were expressed, spin labeled, and reconstituted in DOPC:DOPG lipid bilayers (3:1) and symmetric KCl solutions for spectroscopic measurements.
21. M. Laine, D. M. Papazian, B. Roux, *FEBS Lett.* **564**, 257 (2004).
22. S. R. Durell, Y. Hao, H. R. Guy, *J. Struct. Biol.* **121**, 263 (1998).
23. M. Laine *et al.*, *Neuron* **39**, 467 (2003).
24. Overall architecture, local dynamics, and membrane topology of the voltage sensor were derived from analyses of EPR spectra and power-saturation experiments (*17*–*19*). Motional information was obtained from the inverse of the width of the central resonance line ΔH_0^{-1}, while solvent accessibilities were estimated from power saturation experiments carried out in the presence of either atmospheric oxygen or a water-soluble, neutral NiEdda. Whereas high accessibility to molecular oxygen O_2 (ΠO_2) is indicative of exposure to membrane lipids, residues exposed to the aqueous environment display high NiEdda accessibilities (ΠNiEdda).
25. E. Perozo, L. Cuello, D. Cortes, *Nature Struct. Biol.* **5**, 459 (1998).
26. A. Gross, L. Columbus, K. Hideg, C. Altenbach, W. L. Hubbell, *Biochemistry* **38**, 10324 (1999).
27. Materials and methods are available as supporting material on *Science* Online.
28. W. Kabsch, C. Sander, *Biopolymers* **22**, 2577 (1983).
29. F. Elinder, P. Arhem, H. P. Larsson, *Biophys. J.* **80**, 1802 (2001).
30. D. C. Bell, H. Yao, R. C. Saenger, J. H. Riley, S. A. Siegelbaum, *J. Gen. Physiol.* **123**, 5 (2004).
31. D. M. Starace, F. Bezanilla, *Nature* **427**, 548 (2004).
32. S. A. Monks, D. J. Needleman, C. Miller, *J. Gen. Physiol.* **113**, 415 (1999).
33. K. H. Hong, C. Miller, *J. Gen. Physiol.* **115**, 51 (2000).
34. Y. Li-Smerin, D. H. Hackos, K. J. Swartz, *J. Gen. Physiol.* **115**, 33 (2000).
35. M. F. Sanner, A. J. Olson, J. C. Spehner, *Biopolymers* **38**, 305 (1996).
36. C. C. Huang *et al.*, *Pac. Symp. Biocomput.* **2000**, 230 (2000).

37. We thank H. Mchaourab, R. Nakamoto, C. Ptak, and B. Roux for critically reading the manuscript. F. Bezanilla and K. Swartz provided insightful comments and generated enlightening discussions. H. R. Guy and B. Roux provided coordinates for their K^+ channel models and paddle model. D. Cooper kindly assisted in generating the movie file for the supporting online material. This work was supported in part by NIH.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/491/DC1
Materials and Methods
Figs. S1 to S3
References
Movie S1

14 June 2004; accepted 13 September 2004

Numerical Cognition Without Words: Evidence from Amazonia

Peter Gordon

Members of the Pirahã tribe use a "one-two-many" system of counting. I ask whether speakers of this innumerate language can appreciate larger numerosities without the benefit of words to encode them. This addresses the classic Whorfian question about whether language can determine thought. Results of numerical tasks with varying cognitive demands show that numerical cognition is clearly affected by the lack of a counting system in the language. Performance with quantities greater than three was remarkably poor, but showed a constant coefficient of variation, which is suggestive of an analog estimation process.

Is it possible that there are some concepts that we cannot entertain because of the language that we speak? At issue here is the strongest version of Benjamin Lee Whorf's hypothesis that language can determine the nature and content of thought. The strong version of Whorf's hypothesis goes beyond the weaker claim that linguistic structure simply influences the way that we think about things in our everyday encounters. For example, recent studies suggest that language might affect how people mentally encode spatial relations (*1*–*3*), and how they conceive of the nature of individual objects and their material substances (*4*). However, none of these studies suggest that linguistic structure prevents us from entertaining the concepts that are available to speakers of alternative linguistic systems.

The question of whether linguistic determinism exists in the stronger sense has two parts. The first is whether languages can be incommensurate: Are there terms that exist in one language that cannot be translated into another? The second is whether the lack of such translation precludes the speakers of one language from entertaining concepts that are encoded by the words or grammar of the other language. For many years, the answer to both questions appeared to be negative. Although languages might have different ways in which situations are habitually described, it has generally been accepted that there would always be some way in which one could capture the equivalent meaning in any other language (*5*). Of course, when speaking of translatable concepts, we do not mean terms like "molecule" or "quark," which would not exist in a culture without advanced scientific institutions. Failure to know what molecules or quarks are does not signal an inability to understand the English language—surely people were still speaking English before such terms were introduced. On the other hand, one would question someone's command of English if they did not understand the basic vocabulary and grammar.

Words that indicate numerical quantities are clearly among the basic vocabulary of a language like English. But not all languages contain fully elaborated counting systems. Although no language has been recorded that completely lacks number words, there is a considerable range of counting systems that exists across cultures. Some cultures use a finite number of body parts to count 20 or 30 body tags (*6*). Many cultures use particular body parts like fingers as a recursive base for the count system as in our 10-based system. Finally, there are cultures that base their counting systems on a small number between 2 and 4. Sometimes, the use of a small-number base is recursive and potentially infinite. For example, it is claimed that the Gumulgal South Sea Islanders counted with a recursive binary system: 1, 2, 2'1, 2'2, 2'2'1, and so on (*6*).

The counting system that differs perhaps most from our own is the "one-two-many" system, where quantities beyond two are not counted but are simply referred to as "many." If a culture is limited to such a counting system, is it possible for its members to perceive or conceptualize quantities beyond the limited sets picked out by the counting sequence, or to make what we consider to be quite trivial distinctions such as that between four versus five objects? The Pirahã are such a culture. They live along the banks of the Maici River in the Lowland Amazonia region of Brazil. They maintain a predominantly hunter-gatherer existence and reject assimilation into mainstream Brazilian culture. Almost completely monolingual in their own language, they have a population of less than 200 living in small villages of 10 to 20 people. They have only limited exchanges with outsiders, using primitive pidgin systems for communicating in trading goods without monetary exchange and without the use of Portuguese count words. The Pirahã counting system consists of the words: "hói" (falling tone = "one") and "hoí" (rising tone = "two"). Larger quantities are designated as "baagi" or "aibai" (= "many").

I was able to take three field trips, ranging from 1 week to 2 months, living with the Pirahã along with Daniel Everett and Keren Everett, two linguists who have lived and worked with the tribe for over 20 years and are completely familiar with their language and cultural practices. Observations were informed by their background of continuous and extensive immersion in the Pirahã culture. During my visits, I became interested in the counting system of the Pirahã that I had heard about and wanted to examine whether they really did have only two numbers and how this would affect their ability to perceive numerosities that extended beyond the limited count sequence.

Year 1: Initial observations. On my first week-long trip to the two most up-river Maici villages, I began with informal observations of the Pirahã use of the number words for one and two. I was also interested in the possibility that the one-two-many system might actually be a recursive base-2 system, that their limited number words might be supplemented by more extensive finger counting, or that there might be taboos associated with counting certain kinds of objects as suggested by Zaslavsky in her studies of African counting systems (*7*, *8*). Keren Everett developed some simple tasks to see if our two Pirahã informants could refer to numerosities of arrays of objects using Pirahã terms and any finger counting system they might have. Instructions and interactions with participants were in the Pirahã language. When it was necessary to refer to the numerosity of an array, Keren

Table 1. Use of fingers and number words by Pirahã participant. The arrow (→) indicates a shift from one quantity to the next.

No. of objects	Number word used	No. of fingers
1	hói (= 1)	
2	hoí (= 2)	2
	aibaagi (= many)	
3	hoí (= 2)	3
4	hoí (= 2)	5 → 3
	aibai (= many)	
5	aibaagi (= many)	5
6	aibaagi (= many)	6 → 7
7	hói (= 1)*	1
	aibaagi (= many)	5 → 8
8		5 → 8 → 10
9	aibaagi (= many)	5 → 10
10		5

*This use of "one" might have been a reference to adding one rather than to the whole set of objects.

Department of Biobehavioral Sciences, Columbia University, 525 West 120th Street, New York, NY 10027, USA. E-mail: pgordon@tc.columbia.edu

Everett used the Portuguese number words embedded in Pirahã dialogue. Such terms are understood by the Pirahã to be the language of Brazilians, but their meaning is not understood. In addition to this short session, during the first year trip, I continuously took opportunities to probe for counting abilities in everyday situations.

The outcome of these informal studies revealed the following: (i) There was no recursive use of the count system—the Pirahã never used the count words in combinations like "hói-hoí" to designate larger quantities. (ii) Fingers were used to supplement oral enumeration, but this was highly inaccurate even for small numbers less than five. In addition, "hói" and "hoí," the words for "one" and "two," were not always used to denote those quantities. Whereas the word for "two" always denoted a larger quantity than the word for "one" (when used in the same context), the word for "one" was sometimes used to denote just a small quantity such as two or three or sometimes more. An example of the use of counting words and finger counting is given in Table 1 in one of the informal sessions with an informant who appeared to be in his 50s. Videotaped extracts from the session are included in the supporting online materials (movie S1).

The interpretation of these observations is limited by their informal nature and small sample size. However, the observations are supplemented with 20 years of observation by the Everetts as trained linguists in their analysis of the Pirahã language. One particularly interesting finding is that "hói" appears to designate "roughly one"—or a small quantity whose prototype is one. Most of the time, in the enumeration task, "hói" referred to one, but not always. An analogy might be when we ask for "a couple of *X*s" in English, where the prototypical quantity is two, but we are not upset if we are given three or four objects. However, we surely would be upset if given only one object, because the designation of a single object has a privileged status in our language. There is no concept of "roughly one" in a true integer system. Even the informal use of the indefinite article "a *X*" strictly requires a singular reference. In Pirahã, "hói" can also mean "small," which contrasts with "ogii" (= big), suggesting that the distinction between discrete and continuous quantification is quite fuzzy in the Pirahã language.

Year 2: Experiments in nonverbal numerical reasoning. On my second visit to the Pirahã villages for a 2-month period, I developed a more systematic set of procedures for evaluating the numerical competence of members of the tribe. The experiments were designed to require some combination of cognitive skills such as the need for memory, speed of encoding, and mental-spatial transformations. This would reveal the extent to which such task demands interact with numerical ability, such as it is. Details of the methods are available on *Science* Online (*9*). There were seven participants, who included all six adult males from two villages and one female. Most of the data were collected on four of the men who were consistently available for participation. The tasks were devised to use objects that were available and familiar to the participants (sticks, nuts, and batteries). The results of the tasks, along with schematic diagrams, are presented in Fig. 1. These are roughly ordered in terms of increasing cognitive demand. Any estimation of a person's numerical competence will always be confounded with performance factors of the task. Because this is unavoidable, it makes sense to explore how performance is affected by a range of increasingly demanding tasks.

In the matching tasks (A, B, C, D, and F), I sat across from the participant and with a stick dividing my side from theirs, I presented an array of objects on my side of the stick (below the line in the figures) and they responded by placing a linear array of AA batteries (5.0 cm by 1.4 cm) on their side of the table (above the line). The matching task provides a kind of concrete substitute for counting. It shares the element

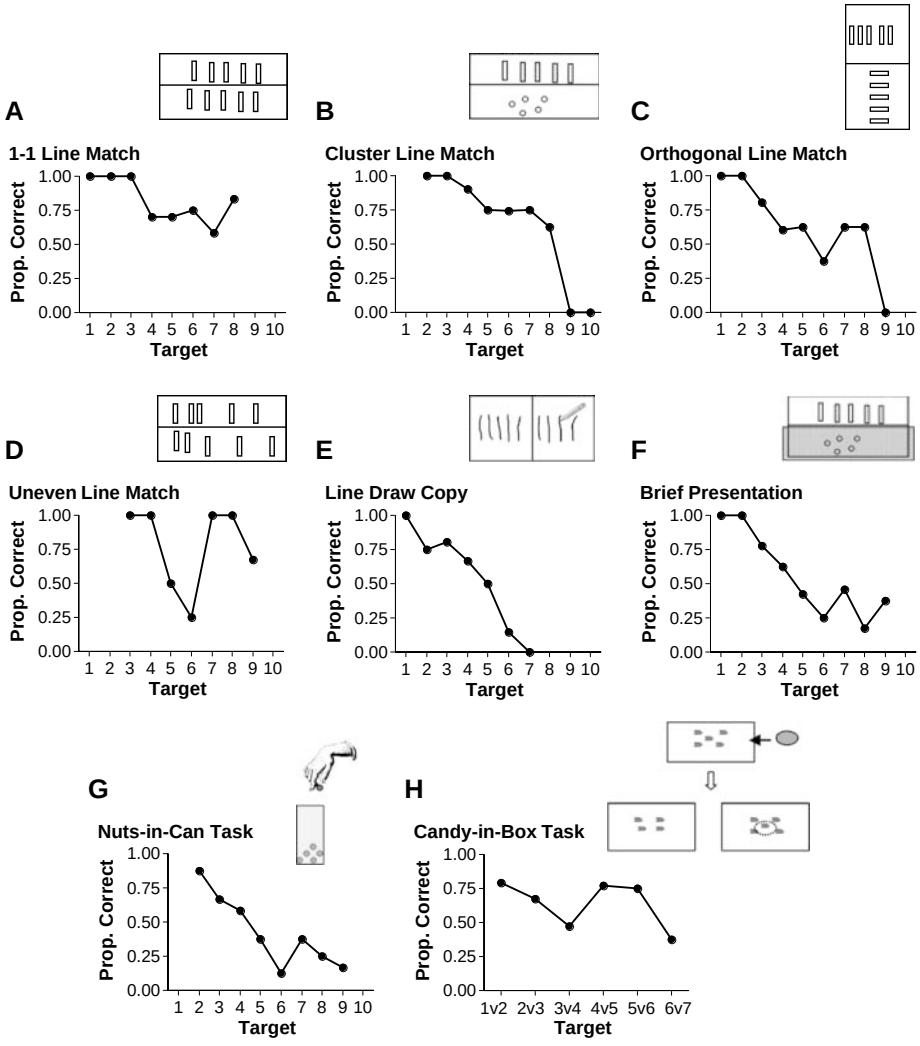


Fig. 1. Results of number tasks with Pirahã villagers (n = 7). Rectangles indicate AA batteries (5.0 cm by 1.4 cm), and circles indicate ground nuts. Center line indicates a stick between the author's example array (below the line) and the participant's attempt to "make it the same" (above the line). Tasks A through D required the participant to match the lower array presented by the author using a line of batteries; task E was similar, but involved the unfamiliar task of copying lines drawn on paper; task F was a matching task where the participant saw the numerical display for only about 1 s before it was hidden behind a screen; task G involved putting nuts into a can and withdrawing them one by one; (participants responded after each withdrawal as to whether the can still contained nuts or was empty); task H involved placing candy inside a box with a number of fish drawn on the lid (this was then hidden and brought out again with another box with one more or one less fish on the lid, and participants had to choose which box contained the candy).

of placing tokens in one-to-one correspondence with individuals in a to-be-counted group. The first matching tasks began with simple linear arrays of batteries. This progressed to clusters of nuts matched to the battery line, orthogonal matching of battery lines, matching of battery lines that were unevenly spaced, and copying lines on a drawing. In all of these matching experiments, participants responded with relatively good accuracy with up to 2 or 3 items, but performance deteriorated considerably beyond that up to 8 to 10 items. In the first simple linear matching task A, performance hovered around 75% up to the largest quantities. Matching tasks with greater cognitive demands required mental transposition of the sample array to the match array without benefit of tagging for numerical quantity. Performance dropped precipitously to 0% for the larger target set sizes in these tasks. One exception was task D with unevenly spaced objects. Although this was designed to be a difficult task, participants showed an anomalous superiority for large numerosities over small. Performance initially deteriorated with increased set size up to 6 items, then shot up to near perfect performance for set size 7 through 10. A likely interpretation of this result was that the uneven spacing for larger set sizes promoted recoding of arrays into smaller configurations of two or three items. This allowed participants to use a chunking strategy of treating each of the subgroups as a matching group.

When time constraints were introduced in task F (exposing the array for only 1 s), performance was drastically affected and there was a clear correlation between set size and accuracy beginning at set size 3. A line-drawing task (E) was highly affected by set size, being one of the worst performances of all. Not only do the Pirahã not count, but they also do not draw. Producing simple straight lines was accomplished only with great effort and concentration, accompanied by heavy sighs and groans. The final two tasks (G and H) required participants to keep track of a numerical quantity through visual displacement. In one case, they were first allowed to inspect an array of nuts for about 8 s. The nuts were placed in a can, and then withdrawn one at a time. Participants were required to say, after each withdrawal, if there were still any nuts left in the can or if it was empty. Performance was predictably strongly affected by set size from the very smallest quantities. The final task involved hiding candy in a box, which had a picture of some number of fish on the lid. The box was then hidden behind the author's back, and two cases were revealed, the original with the candy, and another with one more or one less fish on the lid. For quite small comparisons such as three versus four, performance rarely went over 50% chance responding.

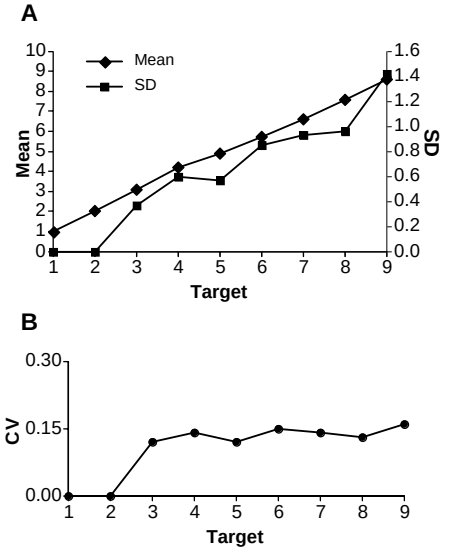


Fig. 2. (A) Mean accuracy and standard deviation of responses in matching tasks and **(B)** coefficient of variation. Figures for individual tasks and individual participants are available in the supporting online materials.

There is a growing consensus in the field of numerical cognition that primitive numerical abilities are of two kinds: First, there is the ability to enumerate accurately small quantities up to about three items, with only minimal processing requirements (*10–16*). I originally termed this ability "parallel individuation" (*17*, *18*), referring to how many items one can encode as discrete unique individuals at the same time in memory. Without overt counting, humans and other animals possess an analog procedure whereby numerical quantities can be estimated with a limited degree of accuracy (*11*, *19–26*). Many researchers believe that large-number estimation, although based on individuated elements, is coalesced into a continuous analog format for mental representation. For example, the discrete elements of a large number array might be represented as a continuous length of a line, where a longer line inexactly represents a larger numerosity.

When people use this analog estimation procedure, the variability of their estimates tends to increase as the target set size increases. The ratio of average error to target set size is known as Weber's fraction and can be indexed by a measure known as the coefficient of variation—the standard deviation of the estimates divided by set size (*23*). Although performance by the Pirahã on the present tasks was quite poor for set sizes above two or three, it was not random. Figure 2 shows the mean response values mapped against the target values for all participants in the simple matching tasks A, B, C, and F. The top graph shows that mean responses and target values are almost identical. This means that the Pirahã participants were trying hard to get the answers correct, and they clearly understood the tasks. The lower graph in Fig. 2 shows that the standard deviation of the estimates increases in proportion to the set size, resulting in a constant coefficient of variation of about 0.15 after set size three, as predicted by the dual model of mental enumeration. This value for the coefficient of variation is about the same as one finds in college students engaged in numerical estimation tasks (*23*). Data for individual tasks and individual participants were consistent with the averaged trends shown in Fig. 2. Graphs are available in the supporting online materials (figs. S2 and S3).

The results of these studies show that the Pirahã's impoverished counting system limits their ability to enumerate exact quantities when set sizes exceed two or three items. For tasks that required additional cognitive processing, performance deteriorated even on set sizes smaller than three. Participants showed evidence of using analog magnitude estimation and, in some cases, they took advantage of spatial chunking to decrease the cognitive demands of larger set sizes. This split between exact enumeration ability for set sizes smaller than three and analog estimation for larger set sizes parallels findings from laboratory experiments with adults who are prevented from explicit counting; studies of numerical abilities in prelinguistic infants, monkeys, birds, and rodents; and in recent studies using brain-imaging techniques (*11*, *23–30*).

The analog estimation abilities exhibited by the Pirahã are a kind of numerical competence that appears to be immune to numerical language deprivation. But because lower animals also exhibit such abilities, robustness in the absence of language is already established. The present experiments allow us to ask whether humans who are not exposed to a number system can represent exact quantities for medium-sized sets of four or five. The answer appears to be negative. The Pirahã inherit just the abilities to exactly enumerate small sets of less than three items if processing factors are not unduly taxing (*31*).

In evaluating the case for linguistic determinism, I suggest that the Pirahã language is incommensurate with languages that have counting systems that enable exact enumeration. Of particular interest is the fact that the Pirahã have no privileged name for the singular quantity. Instead, "hói" meant "roughly one" or "small," which precludes any precise translation of exact numerical terms. The present study represents a rare and perhaps unique case for strong linguistic determinism. The

study also provides a window into how the possibly innate distinction (*26*) between quantifying small versus large sets of objects is relatively unelaborated in a life without number words to capture those exact magnitudes (*32*).

References and Notes

1. S. C. Levinson, *J. Linguist. Anthropol.* **7**, 98 (1997).
2. P. Li, L. Gleitman, *Cognition* **3**, 83 (2002).
3. S. C. Levinson, S. Kita, D. B. M. Haun, B. H. Rasch, *Cognition* **4**, 84 (2002).
4. J. A. Lucy, *Grammatical Categories and Cognition.* (Cambridge Univ. Press, Cambridge, 1992).
5. R. W. Brown, E. H. Lenneberg, *J. Abnorm. Soc. Psychol.* **49**, 454 (1954).
6. K. Menninger, *Number Words and Number Symbols: A Cultural History of Numbers* (MIT Press, Cambridge, MA, 1969).
7. C. Zaslavsky, *Africa Counts: Number and Pattern in African Culture* (Prindle, Weber, and Schmidt, Boston, 1973).
8. R. Gelman, C. R. Gallistel, *The Child's Understanding of Number* (Harvard Univ. Press, Cambridge, MA, 1978).
9. Materials and methods are available as supporting material on *Science* Online.
10. S. Carey, *Mind Lang.* **16**, 37 (2001).
11. L. Feigenson, S. Carey, M. Hauser, *Psychol. Sci.* **13**, 150 (2002).
12. B. J. Scholl, *Cognition* **80**, 1 (2001).
13. T. J. Simon, *Cognit. Dev.* **12**, 349 (1997).
14. L. Trick, Z. W. Pylyshyn, *Psychol. Rev.* **101**, 80 (1994).
15. C. Uller, G. Huntley-Fenner, S. Carey, L. Klatt, *Cognit. Dev.* **14**, 1 (1999).
16. F. Xu, *Cognition* **89**, B15 (2003).
17. P. Gordon, paper presented at the biennial meeting of the Society for Research in Child Development, New Orleans, LA, 25 to 28 March 1993.
18. P. Gordon, paper presented at the European Society for Philosophy and Psychology, Paris, France, 1 to 4 September 1994.
19. W. H. Meck, R. M. Church, *J. Exp. Psychol. Anim. Behav. Processes* **9**, 320 (1983).
20. H. Barth, N. Kanwisher, E. Spelke, *Cognition* **86**, 201 (2003).
21. S. Cordes, R. Gelman, C. R. Gallistel, J. Whalen, *Psychon. Bull. Rev.* **8**, 698 (2001).
22. C. R. Gallistel, *The Organization of Learning* (MIT Press, Cambridge, MA, 1990).
23. J. Whalen, C. R. Gallistel, R. Gelman, *Psychol. Sci.* **10**, 130 (1999).
24. S. Cordes, R. Gelman, C. R. Gallistel, J. Whalen, *Psychon. Bull. Rev.* **8**, 698 (2001).
25. S. Dehaene, *The Number Sense* (Oxford Univ. Press, New York, 1997).
26. B. Butterworth, *What Counts* (Simon & Schuster, New York, 1999).
27. J. S. Lipton, E. S. Spelke, *Psychol. Sci.* **14**, 396 (2003).
28. M. D. Hauser, F. Tsao, P. Garcia, E. S. Spelke, *Proc. R. Soc. London B* **270**, 1441 (2003).
29. S. Dehaene, E. Spelke, P. Pinel, R. Stanescu, S. Tsivkin, *Science* **284**, 970 (1999).
30. J. R. Platt, D. M. Johnson, *Learn. Motiv.* **2**, 386 (1971).
31. Cordes *et al*. (*24*) suggest that analog representations exist even for $n = 2$, because subjects made errors on a task in which counting was suppressed during rapid button pressing. However, errors in this range also occurred when subjects counted and might have been the result of perseveration errors rather than reflecting numerical representations.
32. One can safely rule out that the Pirahã are mentally retarded. Their hunting, spatial, categorization, and linguistic skills are remarkable, and they show no clinical signs of retardation.

33. I thank D. Everett and K. Everett for making this research possible; SIL in Porto Velho and E. Ramos for logistical support; and S. Carey, L. Feigenson, D. Everett, C. Tamis-LeMonda, M. Miozzo, G. Marcus, F. Xu, and K. Adolph for comments on the paper.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1094492/DC1
Methods
SOM Text
Figs. S1 to S3
Movies S1 and S2
References

9 December 2003; accepted 29 July 2004
Published online 19 August 2004;
10.1126/science.1094492
Include this information when citing this paper.

Exact and Approximate Arithmetic in an Amazonian Indigene Group

Pierre Pica,[1] Cathy Lemer,[2] Véronique Izard,[2] Stanislas Dehaene[2]*

Is calculation possible without language? Or is the human ability for arithmetic dependent on the language faculty? To clarify the relation between language and arithmetic, we studied numerical cognition in speakers of Mundurukú, an Amazonian language with a very small lexicon of number words. Although the Mundurukú lack words for numbers beyond 5, they are able to compare and add large approximate numbers that are far beyond their naming range. However, they fail in exact arithmetic with numbers larger than 4 or 5. Our results imply a distinction between a nonverbal system of number approximation and a language-based counting system for exact number and arithmetic.

All science requires mathematics. The knowledge of mathematical things is almost innate in us…. This is the easiest of sciences, a fact which is obvious in that no one's brain rejects it; for laymen and people who are utterly illiterate know how to count and reckon.

Roger Bacon (1214–1294),
English philosopher and scientist

Where does arithmetic come from? For some theorists, the origins of human competence in arithmetic lie in the recursive character of the language faculty (*1*). Chomsky, for instance, stated that "we might think of the human number faculty as essentially an 'abstraction' from human language, preserving the mechanisms of discrete infinity and eliminating the other special features of language" (*2*). Other theorists believe that language is not essential—that humans, like many animals, have a nonverbal "number sense" (*3*), an evolutionarily ancient capacity to process approximate numbers without symbols or language (*4–6*) that provides the conceptual foundation of arithmetic. A third class of theories, while acknowledging the existence of nonverbal representations of numbers, postulates that arithmetic competence is deeply transformed once children acquire a system of number symbols (*7–9*). Language would play an essential role in linking up the various nonverbal representations to create a concept of large exact number (*10–12*).

To elucidate the relations between language and arithmetic, it is necessary to study numerical competence in situations in which the language of numbers is either absent or reduced. In many animal species, as well as in young infants before they acquire number words, behavioral and neurophysiological experiments have revealed the rudiments of arithmetic (*6*, *13–16*). Infants and animals appear to represent only the first three numbers exactly. Beyond this range, they can approximate "numerosity," with a fuzziness that increases linearly with the size of the numbers involved (Weber's law). This finding and the results of other neuroimaging and neuropsychological experiments have yielded a tentative reconciliation of the above theories: Exact arithmetic would require language, whereas approximation would not (*12*, *17–21*). This conclusion, however, has been challenged by a few case studies of adult brain-lesioned or autistic patients in whom language dysfunction did not abolish exact arithmetic; such a finding suggests that in some rare cases, even complex calculation may be performed without words (*22*).

In the final analysis, the debate cannot be settled by studying people who are raised in a culture teeming with spoken and written symbols for numbers. What is needed is a language deprivation experiment, in which neurologically normal adults would be raised

[1]Unité Mixte de Recherche 7023 "Formal Structures of Language," CNRS and Paris VIII University, Paris, France. [2]Unité INSERM 562 "Cognitive Neuroimaging," Service Hospitalier Frédéric Joliot, CEA/DSV, 91401 Orsay Cedex, France.

*To whom correspondence should be addressed. E-mail: dehaene@shfj.cea.fr

without number words or symbols. Although such an experiment is ethically impossible in our Western culture, some languages are intrinsically limited in their ability to express number, sometimes using a very narrow set of number words ("one, two, many") (*23*). These often endangered languages present a rare opportunity to establish the extent and limits of nonverbal arithmetic abilities.

Here, we studied numerical cognition in native speakers of Mundurukú, a language that has number words only for the numbers 1 through 5 (*24*, *25*). Mundurukú is a language of the Tupi family, spoken by about 7000 people living in an autonomous territory in the Pará state of Brazil (Fig. 1). Following regular research stays since 1998, and two pilot studies in 2001 and 2002, one of us (P.P.) traveled through several villages during 2003 and was able to collect data from 55 speakers of Mundurukú in a computerized battery of numerical tests. Ten native speakers of French (mean age 50) served as controls.

The Mundurukú have some contact with nonindigenous culture and individuals, mainly through government institutions and missionaries. Thus, several of them speak some Portuguese, and a few, especially the children, receive some instruction in basic school topics (*26*). To evaluate the potential impact of these variables, we formed two groups of strictly monolingual adults and children without instruction, and we compared their performance with that of more bilingual and educated participants (Fig. 1). Using a solar-powered laptop computer, we collected a large amount of trials in classical arithmetical tasks, including a chronometric comparison test. This allowed us to test whether competence for numbers is present in the absence of a well-developed language for number.

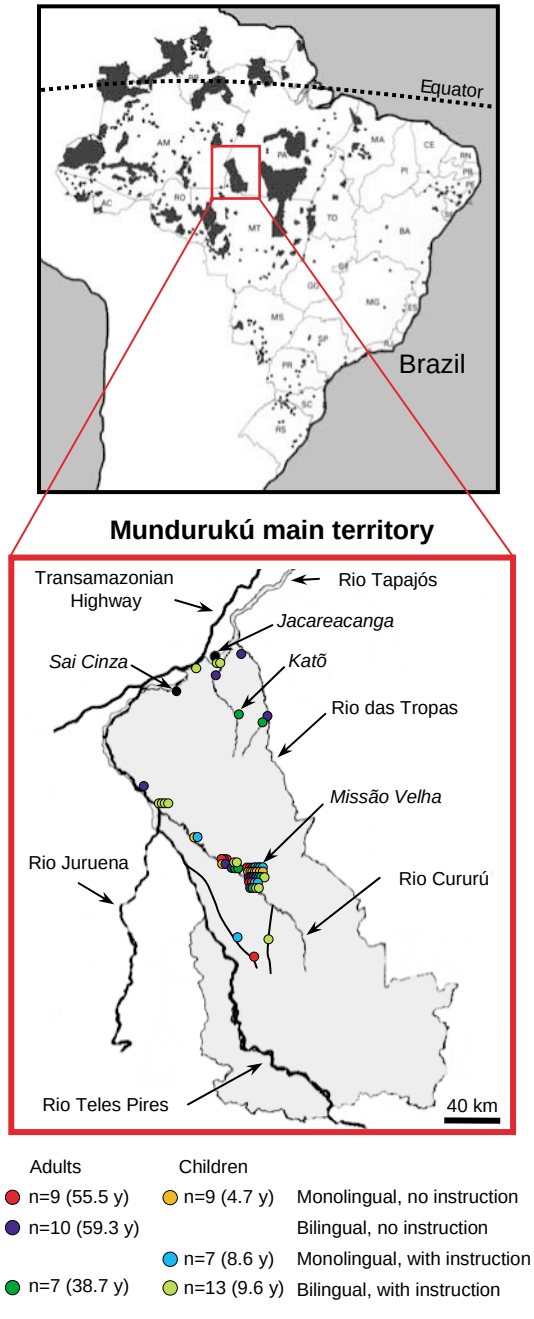


Fig. 1. Location of indigene territories of Brazil (top) and of the main Mundurukú territory where our research was conducted (bottom). Colored dots indicate the villages where participants were tested. The legend at bottom gives the sizes of the six groups of participants and their average age. [Maps adapted with permission from R. Beto, Ed., *Povos indígenas no Brasil* (Instituto Socioambiental, São Paulo, Brazil, 2000), pp. 161, 461].

A first task explored the verbal expressions for numbers in Mundurukú (*26*). Participants were presented with displays of 1 to 15 dots in randomized order, and were asked in their native language to say how many dots were present. This task permitted an objective analysis of the conditions of use of number words. No systematic variation across groups was identified, except for lack of use of the word for "5" in the younger children, and the results were therefore pooled across all groups (Fig. 2). The results confirm that Mundurukú has frozen expressions only for numbers 1 to 5. These expressions are long, often having as many syllables as the corresponding quantity. The words for 3 and 4 are polymorphemic: ebapũg = 2 + 1, ebadipdip = 2 + 1 + 1, where "eba" means "your (two) arms." This possibly reflects an earlier base-2 system common in Tupi languages, but the system is not productive in Mundurukú (expressions such as "eba eba dip" or "eba eba ebapũg" are not used and are judged meaningless).

Above 5, there was little consistency in language use, with no word or expression representing more than 30% of productions to a given target number. Participants relied on approximate quantifiers such as "some" (adesũ), "many" (ade), or "a small quantity" (bũrũmaku). They also used a broad variety of expressions varying in attempted precision, such as "more than one hand," "two hands," "some toes," all the way up to long phrases such as "all the fingers of the hands and then some more" (in response to 13 dots).

The Mundurukú did not use their numerals in a counting sequence, nor to refer to precise quantities. They usually uttered a numeral without counting, although (if asked to do so) some of them could count very slowly and nonverbally by matching their fingers and toes to the set of dots. Our measures confirm that they selected their verbal response on the basis of an apprehension of approximate number rather than on an exact count. With the exception of the words for 1 and 2, all numerals were used in relation to a range of approximate quantities rather than to a precise number (Fig. 2). For instance, the word for 5, which can be translated as "one hand" or "a handful," was used for 5 but also 6, 7, 8, or 9 dots. Conversely, when five dots were presented, the word for 5 was uttered on only 28% of trials, whereas the words for 4 and "few" were each used on about 15% of trials. This response pattern is comparable to the use of

round numbers in Western languages, for instance when we say "10 people" when there are actually 8 or 12. We also noted the occasional use of two-word constructions (e.g., "two-three seeds"), analogous to references to approximate quantities in Western languages (*27*). Thus, the Mundurukú are different from us only in failing to count and in allowing approximate use of number words in the range 3 to 5, where Western numerals usually refer to precise quantities.

If the Mundurukú have a sense of approximate number, they should succeed in approximation tasks with quantities beyond the range for which they have number words. If, however, concepts of numbers emerge only when number words are available, then the Mundurukú would be expected to experience severe difficulties with large numbers. We tested this alternative with the use of two estimation tasks. First, we probed number comparison. Participants were presented with two sets of 20 to 80 dots, controlled for various non-numerical variables (*26*), and were asked to point to the more numerous set (Fig. 3A). Mundurukú participants responded far above chance level in all groups (the minimum was 70.5% correct in the youngest group; all $P < 0.0001$). There was no significant difference among the six Mundurukú groups ($F_{5,46} = 1.50$, $P > 0.20$), which suggests that the small level of bilingualism and instruction achieved by some of the participants did not modify performance. However, average Mundurukú performance was slightly worse than the French controls, thus creating a difference between groups ($F_{6,55} = 2.58$, $P < 0.028$), perhaps due to distraction in some Mundurukú participants (this was the first test that they took).

In literate cultures, number comparison performance is subject to a distance effect: Performance improves as the ratio between the numbers to be compared increases, whether the targets are presented as sets of objects or symbolically as Arabic digits (*28*, *29*). This classical distance effect was also observed in Mundurukú participants: Performance decreased as the ratio varied from 2 to 1.5, 1.3, or 1.2 ($F_{3,138} = 43.2$, $P < 0.0001$). This effect was identical in all groups, including the French controls (group × distance interaction, $F < 1$; see Fig. 3A). Response times were also faster for more distant numbers, in both Mundurukú ($F_{3,90} = 12.9$, $P < 0.0001$) and French participants ($F_{3,26} = 4.93$, $P < 0.008$). Again, although the French controls were globally faster, thus creating a main effect of group ($F_{6,37} = 4.59$, $P < 0.002$), the distance effect was parallel in all groups (interaction $F < 1$). Fitting the performance curve suggested that the Weber fraction, which quantifies the amount of imprecision in number representation (*16*), was 0.17 in Mundurukú, only marginally larger than the value of 0.12 observed in the controls. Thus, the Mundurukú clearly can represent large numbers and understand the concept of relative magnitude (*30*).

We then investigated whether the Mundurukú can perform approximate operations with large numbers. We used a nonsymbolic version of the approximate addition task, which is thought to be independent of language in Western participants (*12*, *17*, *18*). Participants were presented with simple animations illustrating a physical addition of two large sets of dots into a can (Fig. 3B). They had to approximate the result and compare it to a third set. All groups of participants, including monolingual adults and children, performed considerably above chance (minimum 80.7% correct, $P < 0.0001$). Performance was again solely affected by distance ($F_{3,152} = 78.2$, $P < 0.0001$); there was no difference between groups, nor a group × distance interaction (*31*). If anything, performance was higher in this addition + comparison task than in the previous comparison task, perhaps because the operation was represented more concretely by object movement and occlusion. In brief, Mundurukú participants had no difficulty in adding and comparing approximate numbers, with a precision identical to that of the French controls.

Finally, we investigated whether the Mundurukú can manipulate exact numbers. The number sense view predicts that in the absence of spoken or written symbols, number can only be represented approximately, with an internal uncertainty that increases with number (Weber's law). Beyond the range of 3 or 4, this system cannot reliably distinguish an exact number n from its successor $n + 1$. Thus, the Mundurukú should fail with tasks that require manipulation of exact numbers such as "exactly six." To assess this predicted limitation of Mundurukú arithmetic, we used an exact subtraction task. Participants were asked to predict the outcome of a subtraction of a set of dots from an initial set comprising one to eight items (Fig. 3, C and D). The result was always small enough to be named, but the operands could be larger (e.g., 6–4). In the main experiment, for which we report statistics below, participants responded by pointing to the correct result among three alternatives (0, 1, or 2 objects left). The results were also replicated in a second version in which participants named the subtraction result aloud (Fig. 3D).

In both tasks, we observed a fast decrease of performance with the size of the initial number ($F_{7,336} = 44.9$, $P < 0.0001$). This decrease was significant in all Mundurukú groups, although a significant group effect ($F_{5,48} = 3.81$, $P = 0.005$) and a marginal group × size interaction ($F_{35,336} = 1.40$, $P = 0.07$) indicated that performance was slightly better in the more bilingual and educated group, especially when fewer than five dots were present (see Fig. 3D). However, all Mundurukú groups performed much worse

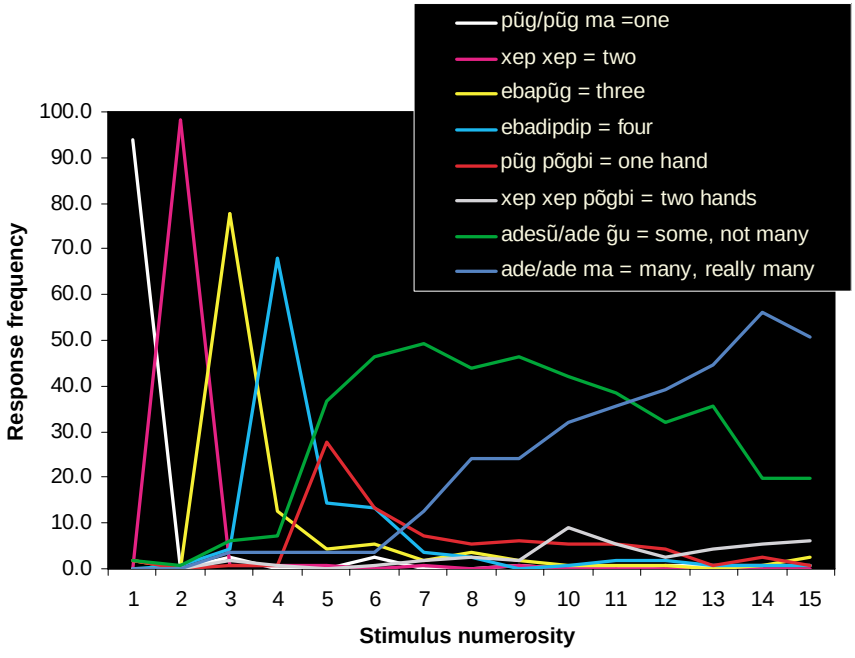


Fig. 2. Number naming in Mundurukú. Participants were shown sets of 1 to 15 dots in random order and were asked to name the quantity. For each quantity on the *x* axis, the graph shows the fraction of times that it was named with a given word or locution. We only present the data for words or locutions produced on more than 2.5% of all trials. For numbers above 5, frequencies do not add up to 100%, because many participants produced rare or idiosyncratic locutions or phrases such as "all of my toes" (a complete list is available from the authors).

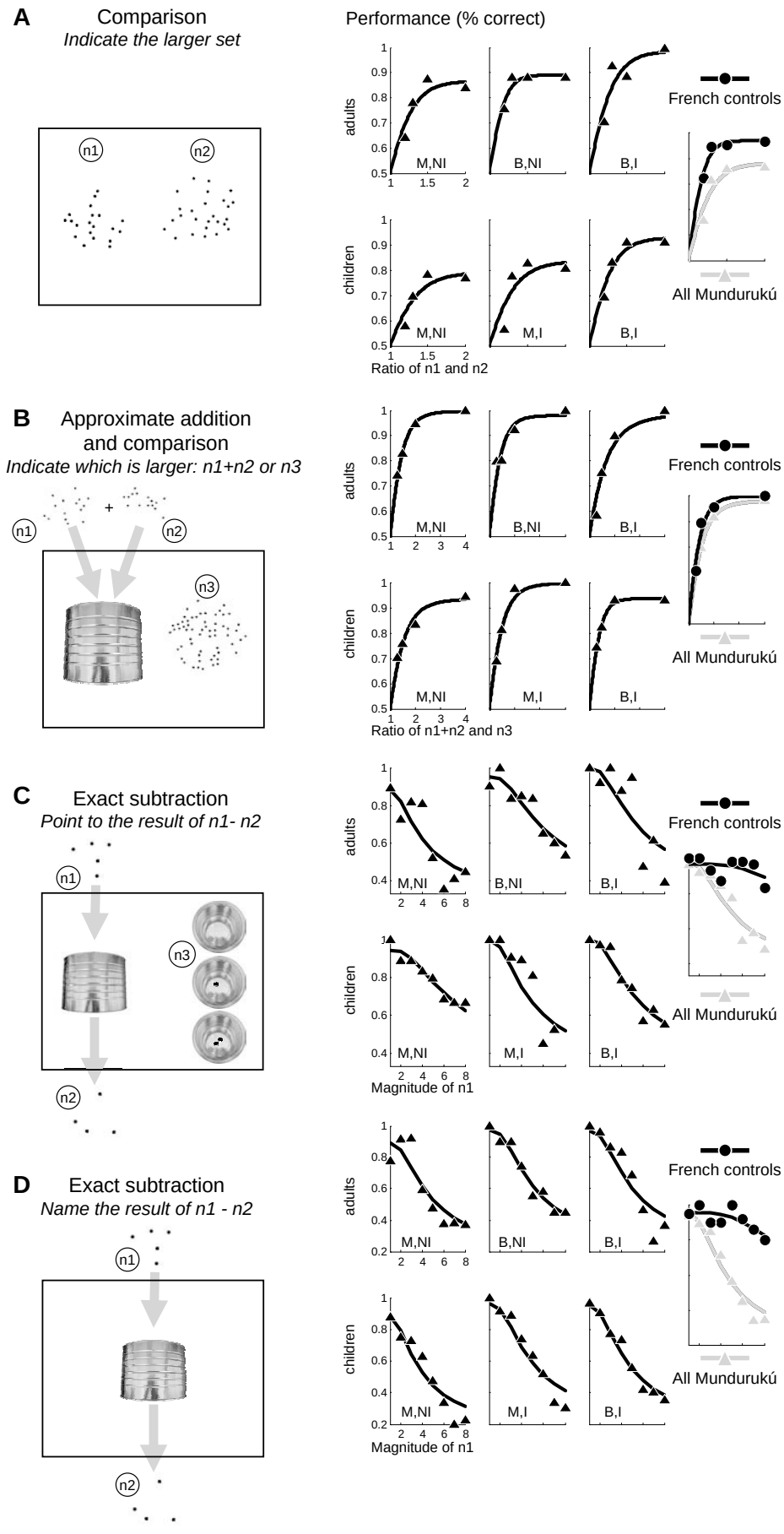


Fig. 3. Performance in four tasks of elementary arithmetic. In each case, the left column illustrates a sample trial (see movie S1). The graphs at right show the fraction of correct trials, in each group separately (M, monolinguals; B, bilinguals; NI, no instruction; I, instruction) as well as averaged across all the Mundurukú and French participants (right graphs). The lowest level on the scale always corresponds to chance performance. For number comparison (**A** and **B**), the relevant variable that determines performance is the distance between the numbers, as measured by the ratio of the larger to the smaller number (e.g., n1/n2 if n1 > n2, n2/n1 otherwise). For exact subtraction (**C** and **D**), the relevant variable is the size of the initial number n1. The fits are based on mathematical equations described in (*26*).

than the French controls, in whom performance was only slightly affected by number size ($F_{7,63}$ = 2.36, P < 0.033). Thus, we observed a highly significant effect of language group (French versus Mundurukú, $F_{1,62}$ = 25.7, P < 0.0001) and a language × size interaction ($F_{7,434}$ = 6.80, P < 0.0001).

The Mundurukú's failure in exact subtraction was not due to misunderstanding of the instructions, because they performed better than chance (indeed, close to 100% correct) when the initial number was below 4. Success within this range might reflect exact verbal coding, or it might reflect a nonverbal parallel individuation of small sets, as also found in preverbal infants (*13*) and nonhuman primates (*14*). Performance also remained above chance for higher values of the initial number (e.g., 49.6% correct for 8 – n problems, chance = 33.3%, P < 0.0001). The entire performance curve over the range 1 to 8 could be fitted by a simple psychophysical equation that supposes an approximate Gaussian encoding of the initial and subtracted quantities, followed by subtraction of those internal magnitudes and classification of the fuzzy outcome into the required response categories (0, 1, or 2). Thus, the Mundurukú still deployed approximate representations, subject to Weber's law, in a task that the French controls easily resolved by exact calculation.

Together, our results shed some light on the issue of the relation between language and arithmetic. They suggest that a basic distinction must be introduced between approximate and exact mental representations of number, as also suggested by earlier behavioral and brain-imaging evidence (*12*, *18*) and by recent research in another Amazon group, the Pirahã (*23*). With approximate quantities, the Mundurukú do not behave qualitatively differently from the French controls. They can mentally represent very large numbers of up to 80 dots, far beyond their naming range, and do not confuse number with other variables such as size and density. They also spontaneously apply concepts of addition, subtraction,

and comparison to these approximate representations. This is true even for monolingual adults and young children who never learned any formal arithmetic. These data add to previous evidence that numerical approximation is a basic competence, independent of language, and available even to preverbal infants and many animal species (*6*, *13*–*16*). We conclude that sophisticated numerical competence can be present in the absence of a well-developed lexicon of number words. This provides an important qualification of Gordon's (*23*) version of Whorf's hypothesis according to which the lexicon of number words drastically limits the ability to entertain abstract number concepts.

What the Mundurukú appear to lack, however, is a procedure for fast apprehension of exact numbers beyond 3 or 4. Our results thus support the hypothesis that language plays a special role in the emergence of exact arithmetic during child development (*9*–*11*). What is the mechanism for this developmental change? It is noteworthy that the Mundurukú have number names up to 5, and yet use them approximately in naming. Thus, the availability of number names, in itself, may not suffice to promote a mental representation of exact number. More crucial, perhaps, is that the Mundurukú do not have a counting routine. Although some have a rudimentary ability to count on their fingers, it is rarely used. By requiring an exact one-to-one pairing of objects with the sequence of numerals, counting may promote a conceptual integration of approximate number representations, discrete object representations, and the verbal code (*10*, *11*). Around the age of 3, Western children exhibit an abrupt change in number processing as they suddenly realize that each count word refers to a precise quantity (*9*). This "crystallization" of discrete numbers out of an initially approximate continuum of numerical magnitudes does not seem to occur in the Mundurukú.

References and Notes

1. J. R. Hurford, *Language and Number* (Blackwell, Oxford, 1987).
2. N. Chomsky, *Language and the Problems of Knowledge* (MIT Press, Cambridge, MA, 1988), p. 169.
3. S. Dehaene, *The Number Sense* (Oxford Univ. Press, New York, 1997).
4. C. R. Gallistel, R. Gelman, *Cognition* **44**, 43 (1992).
5. S. Dehaene, G. Dehaene-Lambertz, L. Cohen, *Trends Neurosci.* **21**, 355 (1998).
6. L. Feigenson, S. Dehaene, E. Spelke, *Trends Cognit. Sci.* **8**, 307 (2004).
7. P. Bloom, *How Children Learn the Meanings of Words* (MIT Press, Cambridge, MA, 2000).
8. H. Wiese, *Numbers, Language, and the Human Mind* (Cambridge Univ. Press, Cambridge, 2003).
9. K. Wynn, *Cognition* **36**, 155 (1990).
10. S. Carey, *Science* **282**, 641 (1998).
11. E. Spelke, S. Tsivkin, in *Language Acquisition and Conceptual Development*, M. Bowerman, S. C. Levinson, Eds. (Cambridge Univ. Press, Cambridge, 2001), pp. 70–100.
12. S. Dehaene, E. Spelke, P. Pinel, R. Stanescu, S. Tsivkin, *Science* **284**, 970 (1999).
13. K. Wynn, *Nature* **358**, 749 (1992).
14. G. M. Sulkowski, M. D. Hauser, *Cognition* **79**, 239 (2001).
15. A. Nieder, E. K. Miller, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7457 (2004).
16. E. M. Brannon, H. S. Terrace, *J. Exp. Psychol. Anim. Behav. Processes* **26**, 31 (2000).
17. E. S. Spelke, S. Tsivkin, *Cognition* **78**, 45 (2001).
18. C. Lemer, S. Dehaene, E. Spelke, L. Cohen, *Neuropsychologia* **41**, 1942 (2003).
19. S. Dehaene, L. Cohen, *Neuropsychologia* **29**, 1045 (1991).
20. H. Barth, N. Kanwisher, E. Spelke, *Cognition* **86**, 201 (2003).
21. J. Whalen, C. R. Gallistel, R. Gelman, *Psychol. Sci.* **10**, 130 (1999).
22. B. Butterworth, *The Mathematical Brain* (Macmillan, London, 1999).
23. P. Gordon, *Science* **306**, 496 (2004); published online 19 August 2004 (10.1126/science.1094492).
24. C. Strömer, *Die sprache der Mundurukú* (Verlag der Internationalen Zeitschrift "Anthropos," Vienna, 1932).
25. M. Crofts, *Aspectos da língua Mundurukú* (Summer Institute of Linguistics, Brasilia, 1985).
26. See supporting data on *Science* Online.
27. T. Pollmann, C. Jansen, *Cognition* **59**, 219 (1996).
28. R. S. Moyer, T. K. Landauer, *Nature* **215**, 1519 (1967).
29. P. B. Buckley, C. B. Gillman, *J. Exp. Psychol.* **103**, 1131 (1974).
30. Comparison performance remained far above chance in two independent sets of trials where the two sets were equalized either on intensive parameters (such as dot size) or on extensive parameters (such as total luminance) [see (*26*)]. Thus, subjects did not base their responses on a single non-numerical parameter. Performance was, however, worse for extensive-matched pairs (88.3% versus 76.3% correct, $P < 0.0001$). We do not know the origins of this effect, but it is likely that, like Western subjects, the Mundurukú estimate number via some simple relation such as the total occupied screen area divided by the average space around the items, which can be subject to various biases [see (*32*)].
31. Performance remained above chance for both intensive-matched and extensive-matched sets (89.5 and 81.8% correct, respectively; both $P < 0.0001$). Although the difference between stimulus sets was again significant ($P < 0.0001$), it was identical in Mundurukú and French subjects. Furthermore, performance was significantly above chance for a vast majority of items (44/51) and was never significantly below chance, making it unlikely that participants were using a simple shortcut other than mental addition. For instance, they did not merely compare n1 with n3 or n2 with n3, because when n1 and n2 were both smaller than n3, they still discerned accurately whether their sum was larger or smaller than the proposed number n3, even when both differed by only 30% (76.3 and 67.4% correct, respectively; both $P < 0.005$).
32. J. Allik, T. Tuulmets, *Percept. Psychophys.* **49**, 303 (1991).

33. This work was developed as part of a larger project on the nature of quantification and functional categories developed jointly with the linguistic section of the Department of Anthropology of the National Museum of Rio de Janeiro and the Unité Mixte de Recherche 7023 of the CNRS, with the agreement of Fundação Nacional do Índio (FUNAI) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) of Brazil. It was supported by INSERM, CNRS, the French Ministry of Foreign Affairs (P.P.), and a McDonnell Foundation centennial fellowship (S.D.). We thank E. Spelke and M. Piazza for discussions, A. Ramos for constant advice, and V. Poxõ, C. Tawé, and F. de Assis for help in testing. Movies illustrating the difficulty of counting for the Mundurukú can be viewed at http://video.rap.prd.fr/videotheques/cnrs/grci.html.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/499/DC1
Materials and Methods
References
Documentary Photos
Movie S1

28 June 2004; accepted 3 September 2004

Separate Neural Systems Value Immediate and Delayed Monetary Rewards

Samuel M. McClure,[1]* David I. Laibson,[2] George Loewenstein,[3] Jonathan D. Cohen[1,4]

When humans are offered the choice between rewards available at different points in time, the relative values of the options are discounted according to their expected delays until delivery. Using functional magnetic resonance imaging, we examined the neural correlates of time discounting while subjects made a series of choices between monetary reward options that varied by delay to delivery. We demonstrate that two separate systems are involved in such decisions. Parts of the limbic system associated with the midbrain dopamine system, including paralimbic cortex, are preferentially activated by decisions involving immediately available rewards. In contrast, regions of the lateral prefrontal cortex and posterior parietal cortex are engaged uniformly by intertemporal choices irrespective of delay. Furthermore, the relative engagement of the two systems is directly associated with subjects' choices, with greater relative fronto-parietal activity when subjects choose longer term options.

In Aesop's classic fable, the ant and the grasshopper are used to illustrate two familiar, but disparate, approaches to human intertemporal decision making. The grasshopper luxuriates during a warm summer day, inattentive to the future. The ant, in contrast,

stores food for the upcoming winter. Human decision makers seem to be torn between an impulse to act like the indulgent grasshopper and an awareness that the patient ant often gets ahead in the long run. An active line of research in both psychology and economics has explored this tension. This research is unified by the idea that consumers behave impatiently today but prefer/plan to act patiently in the future (*1*, *2*). For example, someone offered the choice between $10 today and $11 tomorrow might be tempted to choose the immediate option. However, if asked today to choose between $10 in a year and $11 in a year and a day, the same person is likely to prefer the slightly delayed but larger amount.

Economists and psychologists have theorized about the underlying cause of these dynamically inconsistent choices. It is well accepted that rationality entails treating each moment of delay equally, thereby discounting according to an exponential function (*1*–*3*). Impulsive preference reversals are believed to be indicative of disproportionate valuation of rewards available in the immediate future (*4*–*6*). Some authors have argued that such dynamic inconsistency in preference is driven by a single decision-making system that generates the temporal inconsistency (*7*–*9*), while other authors have argued that the inconsistency is driven by an interaction between two different decision-making systems (*5*, *10*, *11*). We hypothesize that the discrepancy between short-run and long-run preferences reflects the differential activation of distinguishable neural systems. Specifically, we hypothesize that short-run impatience is driven by the limbic system, which responds preferentially to immediate rewards and is less sensitive to the value of future rewards, whereas long-run patience is mediated by the lateral prefrontal cortex and associated structures, which are able to evaluate trade-offs between abstract rewards, including rewards in the more distant future.

A variety of hints in the literature suggest that this might be the case. First, there is the large discrepancy between time discounting in humans and in other species (*12*, *13*). Humans routinely trade off immediate costs/benefits against costs/benefits that are delayed by as much as decades. In contrast, even the most advanced primates, which differ from humans dramatically in the size of their prefrontal cortexes, have not been observed to engage in unprepprogrammed delay of gratification involving more than a few minutes (*12*, *13*). Although some animal behavior appears to weigh trade-offs over longer horizons (e.g., seasonal food storage), such behavior appears invariably to be stereotyped and instinctive, and hence unlike the generalizable nature of human planning. Second, studies of brain damage caused by surgery, accidents, or strokes consistently point to the conclusion that prefrontal damage often leads to behavior that is more heavily influenced by the availability of immediate rewards, as well as failures in the ability to plan (*14*, *15*). Third, a "quasi-hyperbolic" time-discounting function (*16*) that splices together two different discounting functions—one that distinguishes sharply between present and future and another that discounts exponentially and more shallowly—has been found to provide a good fit to experimental data and to shed light on a wide range of behaviors, such as retirement saving, credit-card borrowing, and procrastination (*17*, *18*). However, despite these and many other hints that time discounting may result from distinct processes, little research to date has attempted to directly identify the source of the tension between short-run and long-run preferences.

The quasi-hyperbolic time-discounting function—sometimes referred to as beta-delta preference—was first proposed by Phelps and Pollack (*19*) to model the planning of wealth transfers across generations and applied to the individual's time scale by Elster (*20*) and Laibson (*16*). It posits that the present discounted value of a reward of value u received at delay t is equal to u for $t = 0$ and to $\beta\delta^t u$ for $t > 0$, where $0 < \beta \leq 1$ and $\delta \leq 1$. The β parameter (actually its inverse) represents the special value placed on immediate rewards relative to rewards received at any other point in time. When $\beta < 1$, all future rewards are uniformly downweighted relative to immediate rewards. The δ parameter is simply the discount rate in the standard exponential formula, which treats a given delay equivalently regardless of when it occurs.

Our key hypothesis is that the pattern of behavior that these two parameters summarize—β, which reflects the special weight placed on outcomes that are immediate, and δ, which reflects a more consistent weighting of time periods—stems from the joint influence of distinct neural processes, with β mediated by limbic structures and δ by the lateral prefrontal cortex and associated structures supporting higher cognitive functions.

To test this hypothesis, we measured the brain activity of participants as they made a series of intertemporal choices between early monetary rewards ($\$R$ available at delay d) and later monetary rewards ($\$R'$ available at

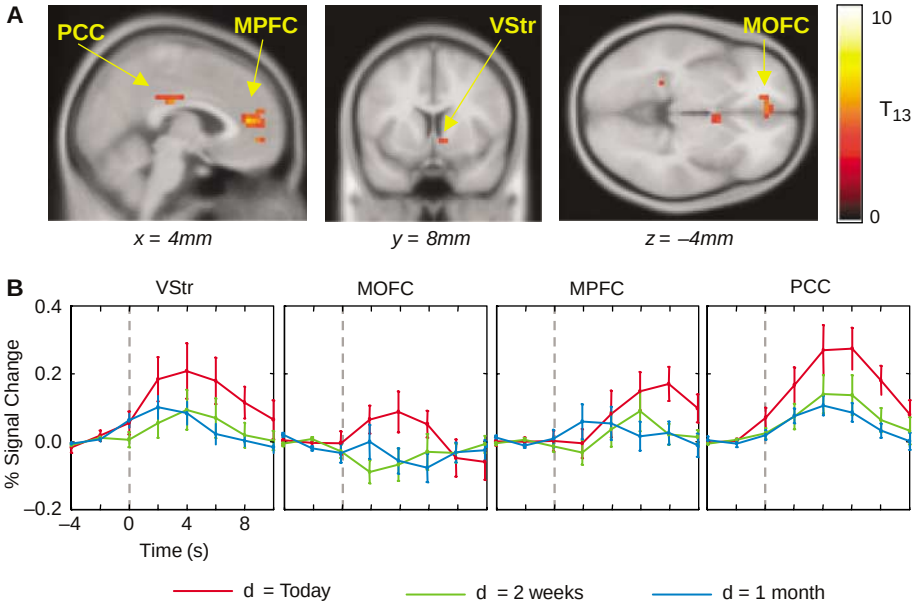


Fig. 1. Brain regions that are preferentially activated for choices in which money is available immediately (β areas). (**A**) A random effects general linear model analysis revealed five regions that are significantly more activated by choices with immediate rewards, implying $d = 0$ (at $P < 0.001$, uncorrected; five contiguous voxels). These regions include the ventral striatum (VStr), medial orbitofrontal cortex (MOFC), medial prefrontal cortex (MPFC), posterior cingulate cortex (PCC), and left posterior hippocampus (table S1). (**B**) Mean event-related time courses of β areas (dashed line indicates the time of choice; error bars are SEM; $n = 14$ subjects). BOLD signal changes in the VStr, MOFC, MPFC, and PCC are all significantly greater when choices involve money available today ($d = 0$, red traces) versus when the earliest choice can be obtained only after a 2-week or 1-month delay ($d = 2$ weeks and $d = 1$ month, green and blue traces, respectively).

[1]Department of Psychology and Center for the Study of Brain, Mind, and Behavior, Princeton University, Princeton, NJ 08544, USA. [2]Department of Economics, Harvard University, and National Bureau of Economic Research, Cambridge, MA 02138, USA. [3]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [4]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15260, USA.

*To whom correspondence should be addressed. E-mail: smcclure@princeton.edu

delay d'; $d' > d$). The early option always had a lower (undiscounted) value than the later option (i.e., $\$R < \$R'$). The two options were separated by a minimum time delay of 2 weeks. In some choice pairs, the early option was available "immediately" (i.e., at the end of the scanning session; $d = 0$). In other choice pairs, even the early option was available only after a delay ($d > 0$).

Our hypotheses led us to make three critical predictions: (i) choice pairs that include a reward today (i.e., $d = 0$) will preferentially engage limbic structures relative to choice pairs that do not include a reward today (i.e., $d > 0$); (ii) lateral prefrontal areas will exhibit similar activity for all choices, as compared with rest, irrespective of reward delay; (iii) trials in which the later reward is selected will be associated with relatively higher levels of lateral prefrontal activation, reflecting the ability of this system to value greater rewards even when they are delayed.

Participants made a series of binary choices between smaller/earlier and larger/later money amounts while their brains were scanned using functional magnetic resonance imaging. The specific amounts (ranging from $5 to $40) and times of availability (ranging from the day of the experiment to 6 weeks later) were varied across choices. At the end of the experiment, one of the participant's choices was randomly selected to count; that is, they received one of the rewards they had selected at the designated time of delivery.

To test our hypotheses, we estimated a general linear model (GLM) using standard regression techniques (*21*). We included two primary regressors in the model, one that modeled decision epochs with an immediacy option in the choice set (the "immediacy" variable) and another that modeled all decision epochs (the "all decisions" variable).

We defined β areas as voxels that loaded on the "immediacy" variable. These are preferentially activated by experimental choices that included an option for a reward today ($d = 0$) as compared with choices involving only delayed outcomes ($d > 0$). As shown in Fig. 1, brain areas disproportionately activated by choices involving an immediate outcome (β areas) include the ventral striatum, medial orbitofrontal cortex, and medial prefrontal cortex. As predicted, these are classic limbic structures and closely associated paralimbic cortical projections. These areas are all also heavily innervated by the midbrain dopamine system and have been shown to be responsive to reward expectation and delivery by the use of direct neuronal recordings in nonhuman species (*22*–*24*) and brain-imaging techniques in humans (*25*–*27*) (Fig. 1). The time courses of activity for these areas are shown in Fig. 1B (*28*, *29*).

We considered voxels that loaded on the "all decisions" variable in our GLM to be candidate δ areas. These were activated by all decision epochs and were not preferentially activated by experimental choices that included an option for a reward today. This criterion identified several areas (Fig. 2), some of which are consistent with our predictions about the δ system (such as lateral prefrontal cortex). However, others (including primary visual and motor cortices) more likely reflect nonspecific aspects of task performance engaged during the decision-making epoch, such as visual processing and motor response. Therefore, we carried out an additional analysis designed to identify areas among these candidate δ regions that were more specifically associated with the decision process.

Specifically, we examined the relationship of activity to decision difficulty, under the assumption that areas involved in decision making would be engaged to a greater degree (and therefore exhibit greater activity) by more difficult decisions (*30*). As expected, the areas of activity observed in visual, premotor, and supplementary motor cortex were not influenced by difficulty, consistent with their role in non–decision-related processes. In contrast, all of the other regions in prefrontal and parietal cortex identified in our initial screen for δ areas showed a significant effect of difficulty, with greater activity associated with more difficult decisions (Fig. 3) (*31*). These findings are consistent with a large number of neurophysiological and neuroimaging studies that have implicated these areas in higher level cognitive functions (*32*, *33*). Furthermore, the areas identified in inferior parietal cortex are similar to those that have been implicated in numerical processing, both in humans and in nonhuman species (*34*). Therefore, our findings are consistent with the hypothesis that lateral prefrontal (and associated parietal) areas are activated by all types of intertemporal choices, not just by those involving immediate rewards.

If this hypothesis is correct, then it makes an additional strong prediction: For choices between immediate and delayed outcomes ($d = 0$), decisions should be determined by the relative activation of the β and δ systems (*35*). More specifically, we assume that when the β system is engaged, it almost always favors the earlier option. Therefore, choices for the later option should reflect a greater influence of the δ system. This implies that choices for the later option should be associated with greater activity in the δ system than in the β system. To test this prediction, we examined activity in β and δ areas for all choices involving the opportunity for a reward today ($d = 0$) to ensure some engagement of the β system. Figure 4 shows that our prediction is confirmed: δ areas were significantly more active than were β areas when participants chose the later option, whereas activity was comparable (with a trend toward greater β-system activity) when participants chose the earlier option.

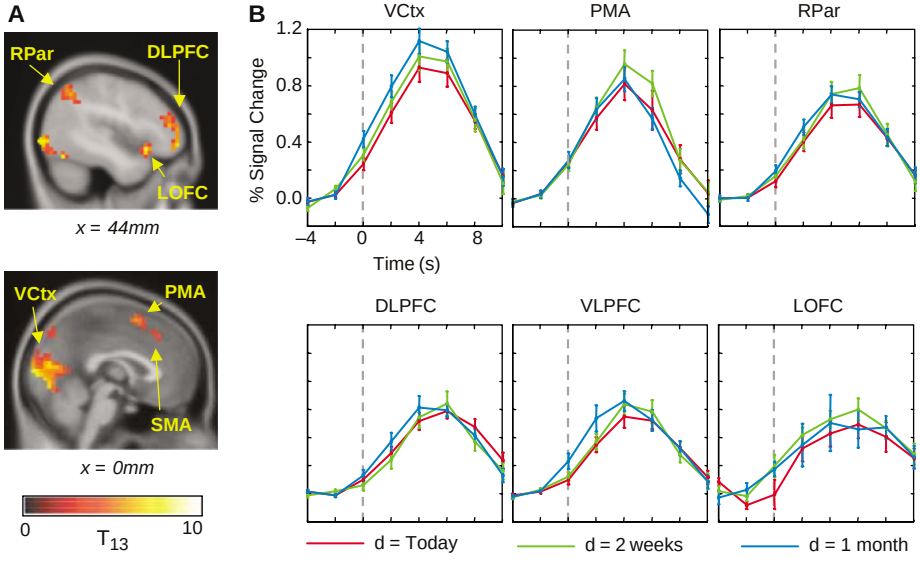


Fig. 2. Brain regions that are active while making choices independent of the delay (d) until the first available reward (δ areas). (**A**) A random effects general linear model analysis revealed eight regions that are uniformly activated by all decision epochs (at $P < 0.001$, uncorrected; five contiguous voxels). These areas include regions of visual cortex (VCtx), premotor area (PMA), and supplementary motor area (SMA). In addition, areas of the right and left intraparietal cortex (RPar, LPar), right dorsolateral prefrontal cortex (DLPFC), right ventrolateral prefrontal cortex (VLPFC), and right lateral orbitofrontal cortex (LOFC) are also activated (table S2). (**B**) Mean event-related time courses for δ areas (dashed line indicates the time of choice; error bars are SEM; $n = 14$ subjects). A three-way analysis of variance indicated that the brain regions identified by this analysis are differentially affected by delay (d) than are those regions identified in Fig. 1 ($P < 0.0001$).

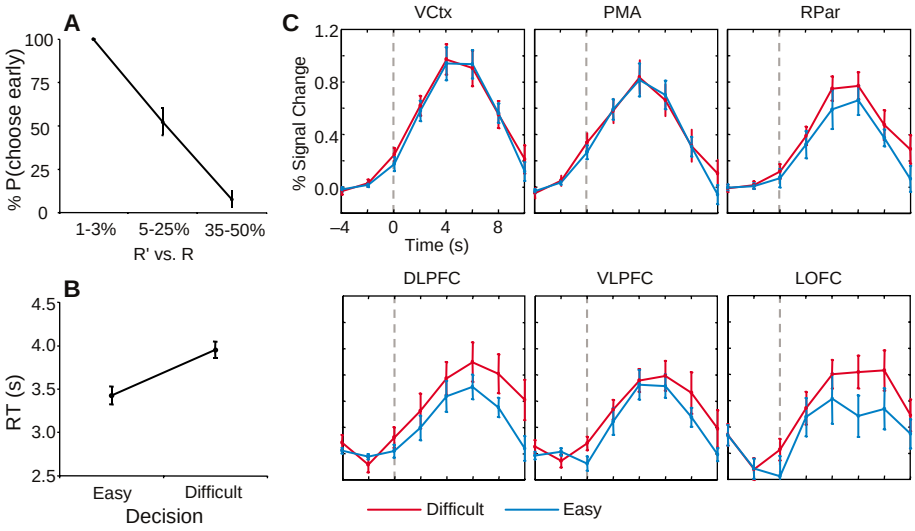


Fig. 3. Differences in brain activity while making easy versus difficult decisions separate δ areas associated with decision making from those associated with non–decision-related aspects of task performance. (**A**) Difficult decisions were defined as those for which the difference in dollar amounts was between 5% and 25%. (**B**) Response times (RT) were significantly longer for difficult choices than for easy choices ($P < 0.005$). (**C**) Difficult choices are associated with greater BOLD signal changes in the DLPFC, VLPFC, LOFC, and inferoparietal cortex (time by difficulty interaction significant at $P < 0.05$ for all areas).

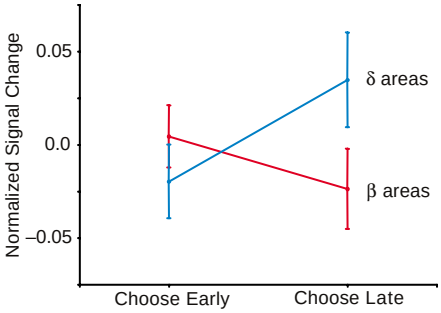


Fig. 4. Greater activity in δ than β areas is associated with the choice of later larger rewards. To assess overall activity among β and δ areas and to make appropriate comparisons, we first normalized the percent signal change (using a *z*-score correction) within each area and each subject, so that the contribution of each brain area was determined relative to its own range of signal variation. Normalized signal change scores were then averaged across areas and subjects separately for the β and δ areas (as identified in Figs. 1 and 2). The average change scores are plotted for each system and each choice outcome. Relative activity in β and δ brain regions correlates with subjects' choices for decisions involving money available today. There was a significant interaction between area and choice ($P < 0.005$), with δ areas showing greater activity when the choice was made for the later option.

In economics, intertemporal choice has long been recognized as a domain in which "the passions" can have large sway in affecting our choices (*36*). Our findings lend support to this intuition. Our analysis shows that the β areas, which are activated disproportionately when choices involve an opportunity for near-term reward, are associated with limbic and paralimbic cortical structures, known to be rich in dopaminergic innervation. These structures have consistently been implicated in impulsive behavior (*37*), and drug addiction is commonly thought to involve disturbances of dopaminergic neurotransmission in these systems (*38*).

Our results help to explain why many factors other than temporal proximity, such as the sight or smell or touch of a desired object, are associated with impulsive behavior. If impatient behavior is driven by limbic activation, it follows that any factor that produces such activation may have effects similar to that of immediacy (*10*). Thus, for example, heroin addicts temporally discount not only heroin but also money more steeply when they are in a drug-craving state (immediately before receiving treatment with an opioid agonist) than when they are not in a drug-craving state (immediately after treatment) (*39*). Immediacy, it seems, may be only one of many factors that, by producing limbic activation, engenders impatience. An important question for future research will be to consider how the steep discounting exhibited by limbic structures in our study of intertemporal preferences relates to the involvement of these structures (and the striatum in particular) in other time-processing tasks, such as interval timing (*40*) and temporal discounting in reinforcement learning paradigms (*41*).

Our analysis shows that the δ areas, which are activated uniformly during all decision epochs, are associated with lateral prefrontal and parietal areas commonly implicated in higher level deliberative processes and cognitive control, including numerical computation (*34*). Such processes are likely to be engaged by the quantitative analysis of economic options and the valuation of future opportunities for reward. The degree of engagement of the δ areas predicts deferral of gratification, consistent with a key role in future planning (*32*, *33*, *42*).

More generally, our present results converge with those of a series of recent imaging studies that have examined the role of limbic structures in valuation and decision making (*26*, *43*, *44*) and interactions between prefrontal cortex and limbic mechanisms in a variety of behavioral contexts, ranging from economic and moral decision making to more visceral responses, such as pain and disgust (*45*–*48*). Collectively, these studies suggest that human behavior is often governed by a competition between lower level, automatic processes that may reflect evolutionary adaptations to particular environments, and the more recently evolved, uniquely human capacity for abstract, domain-general reasoning and future planning. Within the domain of intertemporal choice, the idiosyncrasies of human preferences seem to reflect a competition between the impetuous limbic grasshopper and the provident prefrontal ant within each of us.

References and Notes

1. G. Ainslie, *Psychol. Bull.* **82**, 463 (1975).
2. S. Frederick, G. Loewenstein, T. O'Donoghue, *J. Econ. Lit.* **40**, 351 (2002).
3. T. C. Koopmans, *Econometrica* **32**, 82 (1960).
4. G. Ainslie, *Picoeconomics* (Cambridge Univ. Press, Cambridge, 1992).
5. H. M. Shefrin, R. H. Thaler, *Econ. Inq.* **26**, 609 (1988).
6. R. Benabou, M. Pycia, *Econ. Lett.* **77**, 419 (2002).
7. R. J. Herrnstein, *The Matching Law: Papers in Psychology and Economics*, H. Rachlin, D. I. Laibson, Eds. (Harvard Univ. Press, Cambridge, MA, 1997).
8. H. Rachlin, *The Science of Self-Control* (Harvard Univ. Press, Cambridge, MA, 2000).
9. P. R. Montague, G. S. Berns, *Neuron* **36**, 265 (2002).
10. G. Loewenstein, *Org. Behav. Hum. Decis. Proc.* **65**, 272 (1996).
11. J. Metcalfe, W. Mischel, *Psychol. Rev.* **106**, 3 (1999).
12. H. Rachlin, *Judgment, Decision and Choice: A Cognitive/Behavioral Synthesis* (Freeman, New York, 1989), chap. 7.
13. J. H. Kagel, R. C. Battalio, L. Green, *Economic Choice Theory: An Experimental Analysis of Animal Behavior* (Cambridge Univ. Press, Cambridge, 1995).
14. M. Macmillan, *Brain Cogn.* **19**, 72 (1992).
15. A. Bechara, A. R. Damasio, H. Damasio, S. W. Anderson, *Cognition* **50**, 7 (1994).
16. D. Laibson, *Q. J. Econ.* **112**, 443 (1997).
17. G. Angeletos, D. Laibson, A. Repetto, J. Tobacman, S. Weinberg, *J. Econ. Perspect.* **15**, 47 (2001).
18. T. O'Donoghue, M. Rabin, *Am. Econ. Rev.* **89**, 103 (1999).
19. E. S. Phelps, R. A. Pollak, *Rev. Econ. Stud.* **35**, 185 (1968).
20. J. Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality* (Cambridge Univ. Press, Cambridge, 1979).
21. Materials and methods are available as supporting material on *Science* Online.
22. J. Olds, *Science* **127**, 315 (1958).
23. B. G. Hoebel, *Am. J. Clin. Nutr.* **42**, 1133 (1985).
24. W. Schultz, P. Dayan, P. R. Montague, *Science* **275**, 1593 (1997).

25. H. C. Breiter, B. R. Rosen, *Ann. N.Y. Acad. Sci.* **877**, 523 (1999).
26. B. Knutson, G. W. Fong, C. M. Adams, J. L. Varner, D. Hommer, *Neuroreport* **12**, 3683 (2001).
27. S. M. McClure, G. S. Berns, P. R. Montague, *Neuron* **38**, 339 (2003).
28. Our analysis also identified a region in the dorsal hippocampus as responding preferentially in the d = today condition. However, the mean event-related response in these voxels was qualitatively different from that in the other regions identified by the β analysis (fig. S2). To confirm this, for each area we conducted paired t tests comparing d = today with d = 2 weeks and d = 1 month at each time point after the time of choice. All areas showed at least two time points at which activity was significantly greater for d = today ($P < 0.01$; Bonferroni correction for five comparisons) except the hippocampus, which, by contrast, is not significant for any individual time point. For these reasons, we do not include this region in further analyses. Results are available in (*21*) (fig. S2).
29. One possible explanation for increased activity associated with choice sets that contain immediate rewards is that the discounted value for these choice sets is higher than the discounted value of choice sets that contain only delayed rewards. To rule out this possibility, we estimated discounted value for each choice as the maximum discounted value among the two options. We made the simplifying assumption that subjects maintain a constant weekly discount rate and estimated this value based on expressed preferences (best-fitting value was 7.5% discount rate per week). We then regressed out effects of value from our data with two separate mechanisms. First, we included value as a separate control variable in our baseline GLM model and tested for β and δ effects. Second, we performed a hierarchical analysis in which the effect of value was estimated in a first-stage GLM; this source of variation was then partialed out of the data and the residual data was used to identify β and δ regions in a second-stage GLM. Both of these procedures indicate that value has minimal effects on our results, with all areas of activation remaining significant at $P < 0.001$, uncorrected.
30. Difficulty was assessed by appealing to the variance in preferences indicated by participants. In particular, when the percent difference between dollar amounts of the options in each choice pair was 1% or 3%, subjects invariably opted for the earlier reward, and when the percent difference was 35% or 50%, subjects always selected the later, larger amount. Given this consistency in results, we call these choices "easy." For all other differences, subjects show large variability in preference, and we call these choices "difficult" (Fig. 3A). These designations are further justified by analyzing the mean response time for difficult and easy questions. Subjects required on average 3.95 s to respond to difficult questions and 3.42 s to respond to easy questions (Fig. 3B) ($P < 0.005$). We assume that these differences in response time reflect prolonged decision-making processes for the difficult choices. Based on these designations, we calculated mean blood oxygenation level—dependent (BOLD) responses for easy and difficult choices (Fig. 3C).
31. Because difficulty was associated with longer RT, it was necessary to rule out nonspecific (i.e., non–decision-related) effects of RT as a confound in producing our results. We performed analyses controlling for RT analogous to those performed for discounted value as described above (*29*). This is a conservative test because, as noted above (*30*), we hypothesize that at least some of the variance in RT was related to the decision-making processes of interest. Nevertheless, these analyses indicated that removing the effects of RT does not qualitatively affect our results.
32. E. K. Miller, J. D. Cohen, *Annu. Rev. Neurosci.* **24**, 167 (2001).
33. E. E. Smith, J. Jonides, *Science* **283**, 1657 (1999).
34. S. Dehaene, G. Dehaene-Lambertz, L. Cohen, *Trends Neurosci.* **21**, 355 (1998).
35. This prediction requires only that we assume that activity in each system reflects its overall engagement by the decision and, therefore, its contribution to the outcome. Specifically, it does not require that we assume that the level of activity in either system reflects the value assigned to a particular choice.
36. A. Smith, *Theory of Moral Sentiments* (A. Millar, A. Kinkaid, J. Bell, London and Edinburgh, 1759).
37. J. Biederman, S. V. Faraone, *J. Atten. Disord.* **6**, S1 (2002).
38. G. F. Koob, F. E. Bloom, *Science* **242**, 715 (1988).
39. L. A. Giordano *et al.*, *Psychopharmacology (Berl.)* **163**, 174 (2002).
40. W. H. Meck, A. M. Benson, *Brain Cogn.* **48**, 195 (2002).
41. S. C. Tanaka *et al.*, *Nature Rev. Neurosci.* **7**, 887 (2004).
42. Our results are also consistent with the hypothesis that the fronto-parietal system inhibits the impulse to choose more immediate rewards. However, this hypothesis does not easily account for the fact that this system is recruited even when both rewards are substantially delayed (e.g., 1 month versus 1 month and 2 weeks) and the existence of an impulsive response seems unlikely. Therefore, we favor the hypothesis that fronto-parietal regions may project future benefits (through abstract reasoning or possibly "simulation" with imagery), providing top-down support for responses that favor greater long-term reward and allowing them to compete effectively with limbically mediated responses when these are present.
43. I. Aharon *et al.*, *Neuron* **32**, 537 (2001).
44. B. Seymour *et al.*, *Nature* **429**, 664 (2004).
45. J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, J. D. Cohen, *Science* **293**, 2105 (2001).
46. A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, J. D. Cohen, *Science* **300**, 1755 (2003).
47. T. D. Wager *et al.*, *Science* **303**, 1162 (2004).
48. K. N. Ochsner, S. A. Bunge, J. J. Gross, J. D. Gabrieli, *J. Cogn. Neurosci.* **14**, 1215 (2002).

49. We thank K. D'Ardenne, L. Nystrom, and J. Lee for help with the experiment and J. Schooler for inspiring discussions in the early planning phases of this work. This work was supported by NIH grants MH132804 (J.D.C.), MH065214 (S.M.M.), National Institute on Aging grant AG05842 (D.I.L.), and NSF grant SES-0099025 (D.I.L.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5695/503/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 and S2
References

1 June 2004; accepted 26 August 2004

New Products

MWG Biotech
For more information
+49-8092-8289-929
www.the-mwg.com
www.scienceproductlink.org

Messenger RNA Amplification Kits

The ExpressArt line of messenger RNA (mRNA) amplification kits cover the full range of input total RNA and are available in three versions for nano-, pico-, and microgram amounts of input total RNA. Additional kits for the amplification of bacterial mRNA, for severely degraded RNAs from difficult samples, and from paraffin-embedded tissues will be available soon.

Eppendorf
For more information
+49 40-5 38 01-0
www.eppendorf.com
www.scienceproductlink.org

Probe-Based Assay PCR Mixes

Real-time polymerase chain reaction (PCR) makes use of fluorescent reporter molecules to quantify nucleic acids for assays such as gene expression quantification. Eppendorf has developed a quantitative PCR reaction master mix optimized for probe-based assays. As a result of this optimization, more signal can be generated with the same amount of target DNA, resulting in increased sensitivity. The buffer chemistry used in the master mix continually adjusts the concentration of free magnesium and features a pre-optimized combination of natural and chemically modified deoxynucleotide triphosphates, enhancers, and stabilizers. RealMasterMix is available with or without the passive reference dye Rox.

Schleicher & Schuell
For more information
603-352-3810
www.schleicher-schuell.com
www.scienceproductlink.org

Serum Biomarker Chip

Designed to aid cancer research, the Serum Biomarker Chip is an affordable technology to profile serum biomarkers and discriminate stages of disease progression. The chip enables proteomics researchers to profile and pattern the molecular signature of human serum. The chip contains 120 different antibodies spotted on FAST Slides, a widely accepted protein microarray surface. It is the first high-density antibody chip specific to serum biomarkers related to human cancers of every major organ. The microarray is used for comparing protein abundance profiles of test vs. reference serum samples. Each sample is labeled with two different tags, pooled and probed against the antibody microarray in a competitive-binding fashion. The processed microarray is scanned at two different wavelengths on a two-color fluorescent scanner.

Genevac
For more information
+44-1473-240000
www.genevac.com
www.scienceproductlink.org

High-Throughput Solvent Removal

The HT-12 Series II system is designed to ease solvent evaporation bottlenecks in the drug discovery laboratory. It features a high performance scroll pump that does not require any pumping fluid or lubricant in order to operate. This makes the HT-12 Series suitable for use in harsh chemistry environments where corrosive or volatile solvents are frequently used. The HT-12's high performance and high sample capacities make it suitable for applications that require high throughput evaporation, including combinatorial chemistry, parallel synthesis, liquid chromatography/mass spectrometry purification, compound purification, plate reformatting, and plate replication. The unique design of the multilayer rotor ensures efficient use of valuable laboratory bench space. The HT-12 accommodates a wide range of sample formats. The strong stainless steel rotors are housed in a cast-aluminum vacuum chamber, protected from attack by corrosive vapors by a reinforced perfluorinated coating.

Shimadzu
For more information
800-477-1227
www.ssi.shimadzu.com
www.scienceproductlink.org

High-Throughput HPLC

The LC-2010 high-performance liquid chromatography (HPLC) Series features two models: the LC-2010AHT and the LC-2010CHT with a temperature-controlled sample compartment capable of heating and cooling samples. Both units are fully integrated and include the system controller, membrane degasser, low-pressure gradient unit, ultra-fast autosampler, column oven, and ultraviolet detector with a thermostatted flow cell. The series meets the requirements of many high-throughput applications. Sample injections can be made in as little as 15 s with near-zero sample carryover, superior repeatability, and precision. The micro-volume pump ensures pulse-free solvent delivery and the Dynamic Inlet Valve greatly improves flow rate stability and gradient performance. Sample capacity is provided for 350 × 1 ml vials, 210 × 1.5 ml vials, 100 × 4 ml vials, and 4 standard or deep-well titerplates (96 or 384).

Biometra
For more information
49-551-50686-0
www.biometra.com
www.scienceproductlink.org

Thermocycler

The T3000 Thermocycler features the popular triple block concept, offering three independent blocks in one housing so different protocols can be run at the same time. It is equipped with the latest Peltier technology for better heating and cooling rates and has updated software. With a maximum capacity of 3 × 48 wells, it features high throughput in parallel operation. Three block versions are available for 0.2-ml or 0.5-ml tubes; the combination block can hold both types.

MJ Research
For more information
888-735-8437
www.mjr.com
www.scienceproductlink.org

PCR Enzyme

Phusion High-Fidelity DNA polymerase incorporates a patented DNA-binding element that dramatically increases polymerase processivity. Compared with other "high-fidelity" enzymes for polymerase chain reaction, Phusion polymerase delivers greater fidelity, higher yields per unit, and three- to sixfold shorter reaction times for long targets (1 kb to 28 kb), according to the manufacturer.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and government organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier by visiting www.scienceproductlink.org on the Web, where you can request that the information be sent to you by e-mail, fax, mail, or telephone.